Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

COMPUTATIONAL LINGUISTICS AND TECHNOLOGICAL DEVELOPMENT OF TAMIL

Prof. Rajendran Sankaravelayuthan
Amrita Vishwa Vidyapeetham University
Coimbatore 641 112
rajushush@gmail.com

==========

Abstract

The present monograph entitled "Computational Linguistics" written in Tamil deals mainly with four aspects: computational linguistics in general, the processing of language at different levelvels of analysis such as phonology, morphology, syntax and semantics, applications of computational linguistics and technological development of Tamil. This book is not written at a stretch by me. It is the outcome of my association with computational linguistics for the last twenty five years. So it may read like a compilation.

The book is divided into 13 chapters.

Chapter 1 Introduction: The introduction first introduces the topic of the book 'computational linguistics'. It just explains the importance of studying language though computer and their problems, the handling of language in computer, the tools for language study, the skills needed for linguists to involve themselves in computational linguistics, the skills needed for computer scientist to involve themselves in computational linguistics, natural language processing in Tamil, language technology and development of language technology.

Chapter 2 Natural language Processing: Computer and language processing and the regular expressions and automata are discussed elaborately in this chapter.

Chapter 3 Computational phonology: This chapter describes about four aspects of computational phonology: computer phonology in general, certain important information about acoustic phonetics, linguistics issues in text to speech mechanism and speech to text mechanism.

Chapter 4 Computational morphology: This chapter describes about the computer analysis of morphology. It talks further about the computational modeling of morphology. This is followed by a brief description about morphology of Tamil. The attempts to develop morphological analyzer and morphological generator for Tamil.

Chapter 5 Computational Syntax: This chapter talks about the computational analysis of sentences. Different grammatical formalisms for computational analysis of sentences have been introduced. Shallow parsing is also briefly explained. The aspects of POS tagging, chunking and parsing have been discussed.

Chapter 6 Computational semantics: This chapter talks about linking computational syntax with computational semantics. It also takes about different aspect of computational lexical semantics.

Chapter 7 Spell and grammar checking: This chapter is divided into two sections. The first section talks about the spell checking and the second section talks about grammar checking. In the first section on spell checking, the reasons for the spelling mistakes are discussed first. Then it elaborates on finding and classifying spelling mistakes. This is followed by a discussion on methods of correcting spelling mistakes such as Bayesian rules and Noisy channel models. In the second section on grammar checking, three types of methods are discussed: syntax-based checking, statistics-based checking and rule-based checking.

Chapter 8 Creation of lexical resources: This chapter elaborates on the creation of lexical resources like word indexing, concordance dictionary, machine readable dictionaries, electronic thesaurus, word Net and visual-onto-thesaurus.

Chapter 9 Word sense disambiguation: This chapter discusses at first about lexical ambiguity, different types of lexical ambiguity, a computational model for word sense disambiguation, and different types of word sense disambiguation. This is followed by a discussion on attempts for word sense disambiguation in Tamil language.

Chapter 10 Machine translation: Three major types of machine translation such as human aided machine translation, machine aided human translation and automatic machine translation. This is followed by a discussion on limitation on machine translation and pros and cones of machine translation. After this an elaborate discussion on development of machine translation in India is made.

Chapter 11 Developing computational tools: This chapter briefly talks about developing tools such as machine translation systems, optical character recognition (OCR) and tools for language teaching.

Chapter 12 Technological development of Tamil.

Tamil has initiated its technological development well in advance. It has made use of all the opportunities given to it for making it suitable for digitalization and computerization. The references listed below stand to establish its efforts in fulfilling the need of the day i.e. technological development. Governments, both state and central, funded liberally for the technological development of Tamil. This helped it to develop MT systems, word Net and other NLP systems. Private organizations also contributed for this mission. Many individuals, both from inside and abroad, literally worked for Tamil computing. The organizations such as CIIL, AUKBCRC, Anna University, Amrita University, Tamil Virtual Academy, Tamil University, and Madras University need to appreciated for their efforts in uplifting Tamil in the era of Information Technology. Tamil has switched over to Unicode abandoning other systems. Tamil has comparatively commendable resources and tools for NLP applications. Sumptuous amount of text corpora, speech corpora and parallel corpora are available for Tamil. A good number of speech recognition systems and text to speech systems are developed for Tamil. Reliable morphological analyzers, morphological generators, syntactic parsers, chunkers, shallow parsers, named entity recognition systems optical character recognition system are available for Tamil. Computational semantics also improved in Tamil. There are attempts to develop word sense disambiguation system, question answering system, relationship extraction system, sentiment analysis systems, automatic summarization systems, and co reference resolution systems. Efforts are made to develop text generation systems too for Tamil. Tamil shows only positive symptoms in the technological development.

Chapter 13: Conclusion

Key words: computational linguistics, natural language processing, computational phonology, computational morphology, computational syntax, computational semantics, parsing, analysis, generation, morphological analysis, morphological generation, syntactic parsing, POS tagging, chunking, text-to-speech synthesis, speech to text recognition, spell checking, grammar checking, word sense disambiguation, machine translation, lexical resources, machine readable dictionaries, WordNet, word sense disambiguation, machine translation, computer aided language teaching.

Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

கணினி மொழியியலும் தமிழ்மொழியின் தொழில் நுட்ப வளர்ச்சியும்

(COMPUTATIONAL LINGUISTICS AND TECHNOLOGICAL DEVELOPMENT OF TAMIL)

பேராசிரியர் இராசேந்திரன் சங்கரவேலாயுதன்

அமிர்தா விஷ்வ வித்தயபீடம், கோயம்புத்தூர்

rajushush@gmail.com

கோயம்புத்தூர்

நவம்பர் 2019

என்னுரை

இது என்னால் ஒரே காலகட்டத்தில் எழுதப்பட்ட நூல் அல்ல. இதன் தொடக்கம் 1995இல் என்று எண்ணுகின்றேன். அக்காலகட்டத்தில் தான் பேராசிரியர் சங்கல், பேராசிரியர் வினித் சைத்தன்யா பேராசிரியை அம்பாகுல்கர்னி என்பவர்களின் தொடர்பு எனக்குக் கிடைத்தது. அச்சமயத்தில் கான்பூரில் உள்ள இந்திய தொழில்நுட்ப நிறுவனத்தில் ஆசிரியர்களுக்கும் ஆய்வாளர்களுக்கும் இயற்கைமொழியியல் ஆய்வில் பயிற்சி அளித்து வந்தனர். அத்தகைய பயிற்சியில் தான் நான் மேற் சொன்ன பேராசிரியர்களைச் சந்திக்க எனக்கு வாய்பு கிடைத்தது. முதல் முதலாக கணிப்பொறியைக் கண்ணால் கண்டதும் கையால் தொட்டதும் அங்குதான். அவர்கள் தந்த பயிற்சிதான் என்னை இன்றைய நிலைக்கு ஆளாக்கியது. இந்தி-தமிழ் அனுசாரகா என்ற ஒழுக்கமைப்பு அவர்கள் உதவியால் நான் உருவாக்கி அதன் வெளியீடு கண்டு வியந்து கணினிமொழியியலை என் முன்னேற்றத்தின் கருவியாகத் தேர்ந்தெடுத்தேன். இந்த அறிமுகம் தான் எனக்கு இந்திய அரசின் நிதி நல்கையைப் பெற்றுத் தந்து இந்திய மொழிகளுக்கிடையிலான மொழிபெயர்ப்பு (தமிழ்-மலையாளம் மொழிபெயர்பு), இந்திய மொழிக்கான சொல்வலை உருவாக்கம் (தமிழ்ச் சொல்வலை), இந்திய மொழிகளுக்கான தரவுத்தொகுதி உருவாக்கம் (இந்தி-தமிழ்) என்ற மூன்று ஆய்வுத்திட்டங்களை தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையில் நடத்த வாய்புகளை வாங்கித்தந்தது. 2011 சூனில் தமிழ் பல்கலைக்கழகத்திலிருந்து ஓய்வு பெற்றபின் 2012 சூலையில் கோயம்பத்தூரில் உள்ள அமிர்தா விஷ்வ விஷ்வவித்யபீடம் (அமிர்தா பல்கலைக்கழகம்) என்ற நிறுவத்தில் வருகைதரு பேராசிரியராகப் பணியாற்ற வாய்ப்பு கிடைத்தது. அங்கு நடைபெற்ற இந்திய அரசின் நிதி நல்கையில் நடந்த ஆங்கில-தமிழ் மொழிபெயர்ப்புத் திட்டத்திலும் திராவிடம் மொழிகளின் சொல்வலை என்பதன் ஒரு பகுதியான மலையாளச் சொல்வலை உருவாக்கும் ஆய்வுத்திட்டத்திலும் என் பங்களிப்பைச் செய்ய முடிந்தது. அதைத் தொடர்ந்து சென்னையிலுள்ள தமிழ் இணைய கல்விக்கழகத்தின் நிதி நல்கையின் கீழ் நடைபெற்ற “தமிழ் மொழிக்கான கற்றல் கற்பித்தல் கருவிகள் உருவாக்குதல்” என்ற திட்டத்தின் கீழ் தமிழுக்கான காட்சி-மூலப்பொருண்மையியல் சொற்களஞ்சியத்தை உருவாக்குவதில் பெரும் பங்காற்றினேன். தற்போது இத்தாலியிலுள்ள டிரண்டோ பல்கலைக்கழகத்துடன் இணைந்து சொல்வலையின் நீட்சியாக உலக அளவிலான அறிவு மையம் (Universal Knowledge core) என்ற ஆய்வுத்திட்டதில் ஈடுபட்டுள்ளேன்.

தமிழ்ப் பல்கலைக் கழத்தில் நான் பணியாற்றியபோது மாணவர்களுக்கு கணினி மொழியியல் தொடர்பாக நான் நடத்திய பாடங்களைக் கணினியில் உள்ளீடு செய்துவந்தேன் (கணினி மொழியியல், இயற்கைமொழியியல் ஆய்வு, சொல்வலை, தரவுத்தொகுதி/விரிதரவு மொழியியல், ஒலியியக்கவியலும் பேச்சாய்வும், எழுத்துபிழை மற்றும் இலக்கணப் பிழை திருத்தி என்பன சில). இவையெல்லாம் எனது மடிக்கணினியில் சேகரித்து வைத்து அமிர்தாப் பல்கலைக்கழகத்திற்கு வந்தபின் அவற்றை மேம்படுத்தி academia.edu மற்றும் Research Gate என்ற இணைய தளங்களில் பதிவேற்றி வந்தேன். சில மாதங்களுக்கு முன் பேராசிரியர் திருமலையுடன் மின்னஞ்சல் மூலம் தொடர்புகொண்டு உரையாடிய போது அவரிடம் என்னிடம் மொழியியல் மற்றும் கணினி மொழியியல் தொடர்பான இருபதுக்கும் மேற்பட்ட நூல் வரைவுகள் வெளியிடப்படாமல் இருப்பதைக் கூறினேன். அவர் என்னிடன் அவற்றை நூல்களாக வெளியிடக் கோரியதோடு அல்லாமல் தமது Language in India என்ற மின் திங்களாய்விதழில் வெளியிடவும் உதவிக்கரம் நீட்டினார். இதுவரை அவர் உதவியால் மாதம் ஒரு நூல் என்ற கணக்கில் ஏழு நூல்கள் Language in India-வில் வெளியிடப்பட்டுள்ளன. இது எட்டாவது நூலாகும்.

இது யாருக்காவது பயன்படுமா என்று தெரியவில்லை. சமீபத்தில் பேராசிரியர் சங்கல் அவர்கள் சென்னையில் அண்ணா தொழிநுட்பப் பல்கலைக் கழகத்தில் நடைபேற்ற உலத் தமிழ் இணைய மகாநாட்டில் உரையாற்றியபோது இணையத்தில் இந்திய மொழிகளின் டிஜிட்டல் ஆவணங்கள் மிக அரிதாகும் என்றும் மிக்க குறைந்த விழுக்காடே இந்திய மொழிகளின் பங்களிப்பு என்றும் இந்திய மொழியில் ஏராளம் எழுதி இணையத்தில் பதிவேற்றம் செய்யவேண்டும் என்று அறிவுரை கூறினார். எனவே அதை நிறைவேற்றும் முகமாகவாவது இதை வெளியிடலாம் என்று எண்ணுகின்றேன்.

இந்த நூல் பயனுள்ளதாக இருந்தால் எனக்கு rajushush@gmail.com என்ற மின்னஞ்சல் முகவிரிக்கு எழுதுங்கள். குறை நிறை இருந்தால் சுட்டிக்காட்டுங்கள்

அன்புடன்

இராசேந்திரன் சங்கரவேலாயுதன்

வரிசை எண்	பொருளடக்கம்	பக்கம்
	இயல் 1: அறிமுகம்	25
1.1.	நூலின் நோக்கம்	25
1.2.	முன்னோடியான நூல்கள்	25
1.3.	நூலின் அமைப்பு	26
1.4.	நூலின் பயன்பாடு	27
1.5.	கணிப்பொறிவழி மொழி ஆய்வின் தேவைகளும் சிக்கல்களும்	27
1.6.	கணிப்பொறியில் மொழிகளைக் கையாளுதல்	27
1.7.	மொழி ஆய்விற்குத் தேவைப்படும் கருவிகள்	28
1.8.	மொழி ஆய்வில் மொழியலார்களுக்குத் தேவைப்படும் திறமைகள்	29
1.9.	மொழியாய்வில் கணிப்பொறியாளர்களுக்குத் தேவைப்படும் திறமைகள்	29
1.10.	இயற்கை மொழி ஆய்விற்கான முன்னேற்பாடுகள்	29
1.11.	மொழித் தொழில் நுட்பம்	30
1.12.	தமிழ் மொழித் தொழிநுட்ப வளர்ச்சியின் இன்றைய நிலை	32
1.13.	உலகளவில் கணினி மொழியியலின் தொடக்கமும் வளர்ச்சியும்	35
2.	இயல் 2: கணிபொறிமூலம் மொழி ஆய்வு	36
2.1.	இயற்கை மொழி ஆய்வு	36
2.2.	இயற்கை மொழியும் செயற்கை மொழியும்	36
2.3.	மொழி ஆய்வும் கணிப்பொறியும்	36
2.3.1.	தரவுகளைச் சேகரித்தல்	39
2.3.2.	தகவல்களைப் பெறுதல்	40
2.3.3.	மொழியின் கட்டமைப்பை ஆய்தல்	40
2.3.4.	மொழி நடையை ஆய்தல்	43
2.3.5.	பேச்சைத் தெரிந்து கொள்ளல்	43
2.3.6.	இயந்திர மொழிபெயர்ப்பு	43
2.4.	சீரான வெளிப்பாடுகளும் தானியங்கிகளும்	44
2.4.1.	சீரான வெளிப்பாடுகள்	44
2.4.1.1.	மிக அடிப்படையான சீரான வெளிப்பாட்டு அமைப்பொழுங்கு	45

2.4.1.2.	பிரித்தல், குமுமுதல், முன்னிலை	49
2.4.1.3.	எளிதான எடுத்துக்காட்டுகள்	50
2.4.1.4.	கலவைத்தன்மையான எடுத்துக்காட்டுகள்	50
2.4.1.5.	உயர்நிலை இயக்கிகள்	51
2.4.1.6.	பொதுவான எழுத்துக்களின் குழுமங்களின் மறுகுறியீடுகள்	52
2.4.1.7.	எண்ணுவதற்காகன் சீரான வெளிப்பாட்டு இயக்கிகள்	53
2.4.1.8.	இடச்சாய்கோடு தேவையான சில எழுத்துகள்	53
2.4.1.9.	சீரான வெளிப்பாட்டு இடப்பெயர்ப்பு, நினைவகம், மற்றும் லிசா	53
2.4.2.	முற்றுநிலைத் தானியங்கி	55
2.4.2.1.	முற்றுநிலைத் தானியங்கிகளைக் கொண்டு ஆட்டு மொழியைப் புரிந்துகொள்ளுதல்	56
2.4.2..2	முறையான மொழிகள்	60
2.4.2.3	மற்றொரு எடுத்துக்காட்டு	61
2.4.2.4	உறுதிசெய்யப்படாத முற்றுநிலைத் தானியங்கி	64
2.4.2.5	உறுதிசெய்யப்படாத முற்றுநிலைத் தானியங்கியைக் கோர்வையை ஏற்கப் பயன்படுத்தல்	65
2.4.2.6	தேடலாகத் தெரிந்துகொள்கை	71
2.4.2.6	நிர்ணயிக்கப்பட்ட மற்றும் நிர்ணயிக்கப்படாத தானியங்கிகளை உறவுபடுத்தல்	74
2.4.3.	சீரான மொழிகளும் முற்றுநிலைத் தானியங்கிகளும்	75
2.3.4.	சுருக்கம்	78
	இயல் 3: கணினி ஒலியனியல்	79
3.1.	அறிமுகம்	79
3.2.	ஒலியியல் ஆய்வு	80
3.3.	ஒலிப்பியல் ஆய்வு	82
3.3.1.	ஒலியுறுப்புகள்	82
3.3.2.	மூச்சோட்ட இயக்கம்	87
3.3.3.	ஒலிப்பிகள்	88
3.3.4.	ஒலிப்பிடம்	88

3.3.5	ஒலிப்புமுறை	89
3.3.6.	ஒலிப்பு வருணனை	89
3.3.7.	தமிழ் ஒலிகளின் ஒலிப்பியல் ஆய்வு	94
3.3.7.1.	உயிரொலிகளின் ஒலிப்பியல் ஆய்வு	94
3.3.7.2.	மெய்யொலிகளின் ஒலிப்பியல் ஆய்வு	65
3.4.	ஒலியியக்கவியல் ஆய்வு	96
3.4.1.	தமிழில் ஒலியியக்கவியல் ஆய்வு	103
3.4.2.	தமிழில் ஒரு எடுத்துக்காட்டு ஒலியியக்கவியல் ஆய்வு	104
3.4.2.1	தமிழ் உயிர்களின் ஒலியியக்கவியல் ஆய்வு	106
3.4.2.2.	மெய்யொலிகளின் ஒலியியக்கவியல் ஆய்வு	117
3.5.	உரையிலிருந்து பேச்சாக்கம்	139
3.5.1.	உரையிலிருந்து பேச்சு	140
3.5.2.	ஒலியனியல்	141
3.5.3.	ஒலியன்களும் ஒலியன் விதிகளும்	143
3.5.4.	ஒலியனியல் விதிகளும் முற்றுநிலை மாற்றிகளும்	144
3.5.5.	உரையிலிருந்து பேச்சுக்குவேண்டி உரையை ஒலியன்களுக்குப் பொருத்துதல்	145
3.5.5.1.	உச்சரிப்பு அகராதிகள்	145
3.5.5.2.	அகராதியை நோக்குவதற்கு அப்பாற்பட்ட உரை ஆய்வு	145
3.5.5.3.	முற்றுநிலைமாற்றி அடிப்படையில் உச்சரிப்பு அகராதி	146
3.5.6	உரையிலிருந்து-பேச்சில் மீக்கூறு	148
3.5.6.1	மீக்கூறின் ஒலியல் அல்லது ஒலியியக்கவியல் நோக்குகள்	149
3.5.6.2.	பேச்சு உருவாக்கத்தில் மீக்கூறு	148
3.5.7.	உரையிலிருந்து பேச்சுக்கூட்டிணைப்பாக்கம்	150
3.5.7.1.	பேச்சு	151
3.5.7.2.	பேசு உற்பத்தி	152
3.5.7.3.	பேச்சுக் கேட்புணர்வு	153
3.5.7.4.	பேச்சுக்கூட்டிணைப்பாக்கத்தின் வரலாறு	155
3.5.7.5.	பேச்சுக்கூட்டிணைப்பாக்கத்தின் அணுகுமுறைகள்	159

3.5.7.5.1.	ஒலிப்பியல் சார் கூட்டிணைப்பாகம்	160
3.5.7.5.2.	ஒலிச்செறிவுக் கூட்டிணைப்பாக்கம்	162
3.5.7.5.3.	ஒன்றிணைப்புக் கூட்டிணைப்பாக்கம்	164
3.5.6.5.4.	அலகுத் தேர்வுக் கூட்டிணைப்பக்கம்	166
3.5.7.5.5.	ஒலியுருமைக் கூட்டிணைப்பாக்கம்	167
3.5.7.5.6.	பொருண்மைக்களச் சிறப்புக் கூட்டிணைப்பாக்கம்	168
3.5.7.5.7.	எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பாக்கம்	168
3.5.7.6.	பேச்சுக்கூட்டிணைப்பாக்க நெறிமுறைகளின் நிறைகளும்குறைகளும்	171
3.5..7.	உரையிலிருந்துப் பேச்சாக்க ஒழுங்குமுறைகளின் பயன்பாடுகள்	172
3.5.8.	தமிழில் உரையிலிருந்து பேச்சாக்கம்	174
3.5.8.1.	தமிழில் உரையிலிருந்து பேச்சாக்கத்தின் பின்னடைவு	174
3.5.8.2.	தமிழில் உரையிலிருந்து பேச்சாக்கத்தின் முயற்சிகள்	175
3.5.8.2.1.	பிலிப்ஸ் மற்றும் கூட்டாளிகளின் உரையிலிருந்து பேச்சு மாற்றம்	180
3.5.8.2.2.	உதயகுமார் மற்றும் கூட்டாளிகளின் தமிழுக்கான எழுத்திலிருந்து ஒலியனாக மாற்றும் தீர்மானக் கிளையமைப்புக் கற்றல்	189
3.5.8.2.3.	அருண்குமாரின் தமிழ் உரை-பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறை	196
3.6.	பேச்சிலிருந்து உரையாக்கம்	212
3.6.1.	பேச்சு ஒலிகளும் எழுத்துப்பெயர்ப்பும்	213
3.6.2.	தானியங்கு பேச்சுப் புரிதல்	215
3.6.3.	பேச்சுத் தொழில் நுட்பத்தின் மைல் கற்கள்	218
3.6.4.	பேச்சுச் செயலாக்கத்தின் வகைகள்	220
3.6.5.	பேசு அறிதல் வகைகள்	221
3.6.6.	பேச்சு அறிதல் ஒழுங்குமுறைகளின் பயன்களும் பயன்பாடுகளும்	222
3.6.7.	பொதுவான பேச்சுபுரிதல் ஒழுங்குமுறை	223
3.6.8.	ஒலியியக்க அளபுருவாக்கமும் மாதிரியாக்கமும்	224
3.6.8.1.	ஒலியியக்கப் பண்புக்கூறு ஆய்வு	225
3.6.8.2.	ஒலியியக்க மாதிரிகள்	226
3.6.8.3.	தழுவல்	230

3.6.9.	சொல்சார் மற்றும் உச்சரிப்பு மாதிரியாக்கம்	232
3.6.10.	தமிழில் பேச்சிலிருந்து உரையாக்கம்	235
3.6.10.1.	தமிழில் பேச்சிலிருந்து உரை உருவாக்க முயற்சிகள்	235
3.6.10.2.	பேச்சுத் தரவுகள்	236
3.6.10.3.	தமிழுக்கான அசை அடிப்படையிலான தொடர்ச்சியான பேச்சு அறிவான்	236
3.7.	சுருக்கவுரை	239
	இயல் 4: கணினி உருபனியல்	**241**
4.1.	அறிமுகம்	241
4.2.	உருபனியல் அறிமுகம்	242
4.3.	கணினி உருபனியல் ஆய்வு	253
4.3.1.	உருபனியலை மாதிரிப்படுத்தல்	254
4.3.1.1.	மொழியியல் கோட்பாடுகள்	254
4.3.1.1.1.	சொல்லும் செயற்பாங்கும் மாதிரி	255
4.3.1.1.2.	சொல்லும் அடுக்கும் மாதிரி	255
4.3.1.2.	உளமொழியியல் கோட்பாடு	258
4.3.1.2.1.	மனச்சொற்களஞ்சியத்திலிருந்து உருபனியல் அமைப்பின் சேகரிப்பும் மீளப்பெறுதலும்	258
4.3.1.2.2.	மாதிரி வகைகள்	260
4.3.1.2.1.	தனித்தியங்கும் தேடல் மாதிரி	261
4.3.1.2.2.2.	லோகோகன் மாதிரி	262
4.3.1.2.2.3.	கோகோர்ட் மாதிரி	263
4.3.1.2.2.4.	வகைப்பாடு-பிளவு மாதிரி	263
4.3.1.2.2.5.	பெரிதாக்கப்படவேண்டப்பட்ட உருபனியல்	264
4.3.1.2.2.6.	துணைக்கோள் பதிவுகள் மாதிரிகள்	265
4.3.1.2.2.7.	தனி-பதிவுகள் மாதிரிகள்	265
4.3.1.2.2.8.	எளிய கூறிடப்பட்ட மாதிரி	265
4.3.1.3.	உளமொழியியல் கண்டுபிடித்தங்களின் குறிப்புகள்	265

4.3.1.4.	கணிப்பொறி வழி ஆய்வுக்கான பொருத்தமான மாதிரி	266
4.3.2.	தமிழ் உருபனியல்	267
4.3.2.1.	பெயரும் திரிபும்	268
4.3.2.2.	பெயர்ப்பகுதி	269
4.3.2.3.	திரிபுப் பகுதி	270
4.3.2.3.1.	- த்து என்ற திரிபு ஒட்டுள்ள திரிபுப்பகுதி	271
4.3.2.3.2.	-அற்று என்ற திரிபு ஒட்டுள்ள திரிபுப்பகுதி	271
4.3.2.3.3.	மெய்யொலிகள் இரட்டிப்பால் உருவாக்கப்படும் திரிபுப்பகுதி	272
4.3.2.3.4.	மாற்றிப் பெயர்களின் திரிபுப்பகுதி	273
4.3.2.3.5.	ஒலி நிரவல் சாரிகைகள் இன், அன்	274
4.3.2.4.	எண் குறிப்பு	275
4.3.2.5.	வேற்றுமைக் குறியீடு/உருபு	275
4.3.2.6.	பெயர் அடுக்குகள்	283
4.3.2.6.1.	வேற்றுமை உருபுகளும் சந்தி விதிகளும்	284
4.3.7.	வினை அடுக்குகள்	293
4.4.	தமிழில் உருபனியல் ஆய்விகள்	295
4.4.1.	இராஜேந்திரனின் தமிழ் உருபனியல் பகுப்பாய்வி	297
4.4.2.	கணேசனின் உருபனியல் பகுப்பாய்வி	298
4.4.3.	கபிலனின் தமிழ் வினைகளுக்கான உருபனியல் பகுப்பாய்வி	298
4.4.4.	தெய்வசுந்தரத்தின் உருபனியல் பகுத்துக் குறிப்பான்	298
4.4.5.	AUKBC நிறுவனத் தமிழ் உருபனியல் பகுப்பாய்வி	299
4.4.6.	வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும்	299
4.4.7.	வின்ஸ்டன் குருசின் ஜி.எஸ். மார்ஃப் பகுப்பாய்வியும் உருவாக்கியும்	300
4.4.8.	துரைபாண்டியின் உருபனியல் பகுப்பாய்வி	302
4.4.9.	RCILTS-T-இன் தமிழ் உருபனியல் பகுப்பாய்வி	302
4.4.10.	RCILTS-T –இன் உருபனியல் உருவாக்கி	302
4.4.11.	பொருண்மை மயக்கம் அல்லாத தமிழ் உருபனியல்	302
4.5.	உருபனியலும் முற்றுநிலை மாற்றிகளும்	304

4.5.1.	முற்றுநிலை உருபனியல் பகுத்துக்குறித்தல்	305
4.5.2.	அகராதியும் உருபன் அமைப்பொழுங்கும்	307
4.5.3.	முற்றுநிலை மாறிகளின் உதவியால் உருபனியல் பகுப்பாய்வு	308
4.5.4.	எழுத்தலகு விதிகளும் முற்றுநிலை மாற்றிகளும்	311
4.5.5.	முற்றுநிலை மாற்றியின் அகராதியையும் விதிகளையும் சேர்த்தல்	311
4.5.6	பிசிகிகிமோ	312
4.6.	சுருக்கம்	314
5.	**இயல் 5: கணினித் தொடரியல்**	**315**
5.1.	அறிமுகம்	315
5.2.	கணிப்பொறி வழி மொழி ஆய்வுக்கான இலக்கண வடிவமைப்புகள்	317
5.2.1.	சூழல்வரையிலா இலக்கணம்	315
5.2.1.1.	அண்மையுறுப்பு அணுகுமுறை தொடரமைப்புச் சட்டகம்	317
5.2.1.2.	பகுத்தாயும் நடைமுறைத்திறன்	323
5.2.1.2.1.	மேலிருந்து கீழ் மற்றும் கீழிருந்து மேல் பகுத்தாய்த்தல்	324
5.2.1.2.2.	ஆழம் முதல் மற்றும் அகலம் முதல் பகுத்தாய்த்தல்	328
5.2.1.2.3.	நிலைமாற்ற வலைப்பின்னலால் பகுத்தாய்த்தல்	330
5.2.1.2.3.1.	மறுதரவு நிலைமாற்ற வலைப்பின்னல்கள்	332
5.2.1.2.3.2.	நிலைபெறு பெரிதாக்கப்பட்ட நிலைமாற்ற வலைப்பின்னல்கள்	332
5.2.2.	பாணினி இலக்கண வடிவவாதம்	338
5.2.2.1.	பாணினி பகுத்தாய்வான்	343
5.2.2.1.1.	முக்கியப் பகுத்தாய்வான்	343
5.2.2.1.2.	குறிப்பிட்ட இடம்சார்ந்த சொற்களைக் குழுமுதல்	344
5.2.2.2	பகுத்தாய்வின் முக்கியத்துவம்	345
5.2.3.	சார்புப் பகுப்பாய்வு	345
5.3.	ஆழமில்லாப் பகுப்பாய்வு	347
5.4.	சொல்வகைப்பாட்டு அடையாளக் குழுமங்களும் சொல்வகைப்பாடு அடையாளப்படுத்துகையும்	348
5.4.1.	சொல்வகைப்பாட்டு அடையாளக்குழுமங்கள் கண்டுபிடிக்கும்	349

	முயற்சிகள்	
5.4.2.	சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பு	351
5.4.3.	இந்தியமொழிகளின் மைய நிறுவனத்தின் சொல்வகைப்பாடு அடையாளக் குழுமம்	353
5.4.4	எ.யு.கே.பி.சி நிறுவனத்தின் சொல்வகைப்பாடு அடையாளக் குழுமம்	355
5.4.5.	ஹைதராபாத் IIITஆல் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல்வகைப்பாடு அடையாளக் குழுமம்	360
5.4.6.	அமிர்தா பல்கலைக்கழகத்தின் சொல்வகைப்பாட்டு அடையாளக் குழுமம்	363
5.4.7.	தமிழுக்கான பிற சொல்வகைபாட்டு அடையாளப்படுத்திகள்	363
5.4.7.1.	வாசு அரங்கநாதனின் டேக்தமிழ்	363
5.4.7.2.	கணேசனின் சொல்வகைப்பாட்டு அடையாளப்படுத்தி	364
5.4.7.3	RCILTS-இன் தமிழ் கதம்பம்	364
5.4.7.4.	பாஸ்கரன் மற்றும் பிறரின் இந்திய மொழிக்கான பொதுச் சொல்வகைப்பாட்டு அடையாளக்குழுமச் சட்டகம்	354
5.4.7.5.	இந்தியமொழிகளுக்கான பிஸ் சொல்வகைப்பாட்டு அடையாளக்குழுமம்	371
5.4.7.6.	சர்வேஸ்வரன் மற்றும் மகேசன் என்போரின் விதி அடிப்படையிலான படிநிலை அடையாளக் குழுமம்	373
5.4.8.	மூவகை சொல்வகை அடையாளப்படுத்தும் வழிமுறை வரைவுகள்	373
5.4.8.1.	விதி அடிப்படையிலான சொல்வகைப்பாட்டு அடையாளப்படுத்தல்	374
5.4.8.2.	புள்ளியியல் அடிப்படையிலான சொல்வகைப்பாட்டு அடையாளப்படுத்தல்	375
5.4.8.3.	மாற்றம் அடிப்படையிலான சொல்வகைப்பாட்டு அடையாளப்படுத்தல்	376
5.4.8.3.1.	எவ்வாறு மாற்றம் அடிப்படையிலான சொல்வகைப்பாட்டு அடையாளப்படுத்தல் விதிகள் பயன்படுத்தப்படுகின்றன	377
5.4.8.3.2.	எவ்வாறு மாற்றம் அடிப்படையிலான சொல்வகைப்பாட்டு	377

	அடையாளப்படுத்தல் விதிகள் கற்கப்படுகின்றன	
5.5.	தொடரியல் பகுப்பான்கள்	378
5.5.1.	தமிழில் தொடரியல் பகுத்தாய்வான்கள்	380
5.5.1.1.	பாஸ்கரனின் தொடரியல் பகுத்தாய்வான்	380
5.5.1.2.	எயுகேபிசி ஆய்வுமையத்தின் பெயர்த்தொடர் பகுப்பான்	380
5.5.1.3.	RCILTS –இன் வானவில்	381
5.5.1.4.	குமார சண்முகத்தின் தொடரியல் பகுத்தாய்வான்	382
5.5.2.	இந்தியமொழிகளுக்கிடையிலான மொழிபெயர்ப்புக்கான தொடரியல் பகுத்தாய்வான்	382
5.6.	தமிழில் மொழிநடை ஆய்வி	383
5.7.	சுருக்கவுரை	383
6	இயல் **7:** கணினிப் பொருண்மையியல்	384
6.1.	அறிமுகம்	384
6.2.	பொருண்மையை உருப்படுத்தம் செய்தல்	384
6.3.	பொருண்மையியல் ஆய்வு	384
6.4.	தொடரியலால் இயக்கப்படும் பொருண்மையியல்	384
6.5.	சூழல் கட்டுப்பாடில்லாத விதிகளில் பொருண்மையின் விரிவாக்கங்கள்	385
6.6.	முந்தையப் பகுப்பானில் பொருண்மையியல் ஆய்வை ஒருங்கிணைத்தல்	385
6.7.	கணினிச் சொற்பொருண்மையியல்	385
6.7.1.	மொழியியல் மற்றும் உளமொழியல் பார்வையில் சொற்பொருண்மையியல்	386
6.7.1.1.	சொற்பொருண்மையியல் ஆய்வில் உளமொழியியலாரின் பங்கு	387
6.7.1.1.1.	காலின் மற்றும் குல்லியன் மாதிரி	387
6.7.1.1.2.	மில்லர் மற்றும் ஜாண்சன்-லைட் மாதிரி	388
6.7.1.1.3.	புரிதல் சொல்சார் அணுகல்	390
6.7.1.1.4.	உருவாக்கலில் சொல்சார் அணுகல்	390
6.7.1.1.5.	தானியக்க அல்லது கட்டுப்பட்ட செயற்பாங்கா? இணையான	391

	அல்லது தொடரான நடைமுறை	
6.7.1.1.6.	கூறாய்வுப்பொருண்மையியல்	391
6.7.1.1.7.	செயல்முறைப் பொருண்மையியல்	393
6.7.1.2.	மொழியியல் சொற்பொருண்மையியல்	394
6.7.1.2.1.	பங்கேற்பாளர் அமைப்பு	394
6.7.1.2.2.	மைப்பொருள் பங்களிப்புகள்	395
6.7.1.2.3.	தேர்வுக் கட்டுப்பாடுகள்	397
6.7.1.2.4.	சொற்பொருண்மை உறவுகள்	398
6.7.1.2.5.	மூலப்பொருண்மையியல்கள்	404
6.7.1.2.6.	சொல்சார் கருத்துரு அமைப்பு	406
6.7.1.2.7.	நிகழ்வு அமைப்புகள் மற்றும் காலச் சொல்சார் தகவல்	407
6.7.1.2.8.	குண மற்றும் மரபுரிமை அமைப்புகள்	407
6.7.1.2.9.	பொருண்மை-பனுவல் கோட்பாட்டு சொல்சார் செயல்பாடுகள்	409
6.7.2.	மன அகரதியும் இயந்திர அகராதியும்	410
6.7.2.1.	இயற்கை மற்றும் செயற்கைப் பொருண்மை அலகுகளின் பொதுவான பண்புகள்	411
6.7.2.2.	நடைமுறைப்படுத்தும் நுட்பத்தில் உள்ள அடிப்படை வேறுபாடுகள்	411
6.7.3.	சொற் பொருண்மையியலிலிருந்து பனுவல் ஆய்வு	412
6.7.4.	சொற்பொருண்மையியல் அகராதியா கலைக்களஞ்சியமா	414
6.8.	சுருக்கவுரை	416
7.	**இயல் 7: எழுத்துப்பிழைத் திருத்தமும் இலக்கணப் பிழைதிருத்தமும்**	417
7.1.	அறிமுகம்	417
7.2.	எழுத்துப்பிழை திருத்தி	418
7.2.1.	எழுத்துப் பிழை நிகழ்வதற்கான வாய்ப்புகள்	418
7.2.2.	எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் சிக்கல்களின் வகைகள்	419
7.2.3.	மொழியில் இல்லாத சொற்களைக் கண்டறிந்து மாற்றுதலுக்கான செயல்திட்டம்	420
7.2.4.	பிழை நீக்கும் வழிமுறைகள்	421
7.2.4.1.	பெசியன் முறை	421

7.2.4.2.	குறைந்த திருத்து தூரம்	425
7.3.	இலக்கணப்பிழை திருத்தி	427
7.3.1.	தொடரியல் அடிப்படையிலான திருத்தம்	427
7.3.2.	புள்ளியியல் அடிப்படையிலான திருத்தம்	428
7.3.3.	விதி அடிப்படையிலான திருத்தம்	432
7.4.	நடை மற்றும் இலக்கணத் திருத்தி	432
7.4.1.	கோட்பாட்டுப் பின்னணி	433
7.5.	தமிழில் எழுத்துப்பிழை மற்றும் இலக்கணப்பிழைத் திருத்தி உருவாக்கம்	435
7.5.1.	தமிழ்ப் பல்கலைக்கழகப் பிழை திருத்தி	436
7.5.2.	தனபாலன் மற்றும் பிறரின் தமிழ் எழுத்துப்பிழை திருத்தி	436
7.5.2.1.	தமிழ்ப் பிழை திருத்தத்தின் சிக்கல்கள்	436
7.5.2.2.	தமிழ்ப் பிழை திருத்தியின் தொகுதிகள்	437
7.5.2.2.1.	பெயர்களைச் சரிபார்த்தல்	438
7.5.2.2.2.	வினைச்சொற்களைச் சரிபார்த்தல்	440
7.5.2.2.3.	பின்னுருபுகளைச் சரிபார்த்தல்	442
7.5.2.2.4.	ஒன்றுபோல் உச்சரிக்கப்படும் எழுத்துகள்	442
7.5.2.2.5.	அண்மை விசைப் பிழைகள்	443
7.5.2.2.6.	எழுத்து நிலைநிறுத்து விதிகள்	444
7.5.2.2.7.	அகராதியைப் பார்த்தல்	444
7.5.2.2.8.	பரிந்துரை உருவாக்கம்	444
7.5.2.3.	முடிவுரை	445
7.5.3.	விஜய் சங்கர் ராமின் தமிழ்ப் பிழைதிருத்தி உருவாக்குதல்	445
7.5.3.1.	எழுத்துப் பிழைதிருத்தலின் பல்வேறு அணுகுமுறைகள்	445
7.5.3.2.	இந்திய மொழிகளுக்கு இருக்கின்ற எழுத்துப் பிழை திருத்திகள்	446
7.5.3.3.	பிழை கண்டுபிடித்தல்	446
7.5.3.3.1.	தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை	447
7.5.3.3.2.	மொழியியல் அடிப்படையிலான அணுகுமுறை	449

7.5.4.	தெய்வசுந்தரத்தின் மென் தமிழ் சொல்லாளர்	453
7.5.5.	எடுத்துக்காட்டான பிழை திருத்தி	454
7.6.	சுருக்கம்	456
8	**இயல் 8: சொல்மூலங்களை உருவக்குதல்**	461
8.1.	அறிமுகம்	461
8.2.	சொல்லடைவு உருவாக்குதல்	461
8.3.	சொல்லாட்சி அகராதி மற்றும் சொற்பொருளடைவு உருவாக்குதல்	462
8.4.	அகராதி உருவாக்குதல்	463
8.4.1.	தமிழில் கணிப்பொறி வழி அகராதி உருவாக்குதல்	463
8.4.2.	அகராதி உருவாக்கும் படிநிலைகள்	463
8.4.3.	கணிப்பொறி வழி அகராதி-வரலாற்றுச் சுருக்கம்	464
8.4.4.	கணிப்பொறி வழி அகராதி: முறைமைகள்	464
8.4.5.	அகராதி உருவாக்குவதில் இயற்மொழி ஆய்வு	465
8.4.6.	முடிவுரை	466
8.5.	சொற்களஞ்சியம் உருவாக்குதல்	467
8.6.	சொல்வலை உருவாக்குதல்	468
8.6.1.	இந்தோ சொல்வலை	470
8.6.2.	திராவிடச் சொல்வலை	470
8.6.3.	யூரோ சொல்வலை	471
8.6.4.	தமிழ்ச் சொல்வலை	471
8.6.4.1.	அகராதியும் சொல்வலையும்	471
8.6.4.2.	சொற்களஞ்சியமும் சொல்வலையும்	472
8.6.4.3.	சொல்வலையில் உறவுகள்	472
8.6.4.4.	பெயர்ச் சொல்வலை	473
8.6.4.4.1.	சொற்களின் படிநிலை அமைப்பு	473
8.6.4.4.2.	தனித்தனியான தொடக்கிகள்	474
8.6.4.4.3.	வேறுபடுத்தும் பொருண்மைக் கூறுகள்	475
8.6.4.4.4.	சினை-முழு உறவு	476

8.6.4.4.5.	பெயர்ச் சொல்வலையின் பொருண்மை உறவுகளின் சுருக்கம்	476
8.6.4.5.	வினைச் சொல்வலை	478
8.6.4.5.1.	வினைச்சொற்களைச் சொற்களங்களாகப் பிரித்தல்	478
8.6.4.5.2.	வினைகளுக்குத் தனியான தொடக்கிகள்	478
8.6.4.5.3.	ஒருபொருள்பன்மொழியக் குழுமங்களாக வினைகள்	478
8.6.4.5.4.	வினைகளின் பொருண்மைக் கூறாய்வு	478
8.6.4.5.5.	வினைகளுக்கிடையிலான சொல் மற்றும் பொருண்மை உறவுகள்	479
8.6.4.5.6.	வினைகளுக்கிடையிலான பல்பொருள் ஒருமொழியம்	479
8.6.4.5.7.	தொடரியல் பண்புகளும் பொருண்மை உறவுகளும்	479
8.6.4.5.8.	வினைச் சொல்வலையின் பொருண்மை உறவுகளின் சுருக்கம்	479
8.6.4.6.	பெயரடை மற்றும் வினையடைச் சொல்வலை	480
8.6.4.7.	சொல்வலையைத் திட்டமிட்டு நடைமுறைப் படுத்தல்	481
8.6.5.	முடிவுரை	482
8.7.	காட்சி மூலப்பொருண்மையியல் சொற்களஞ்சிய உருவாக்கம்	482
8.8.	சுருக்க உரை	494
9.	**இயல் 9: சொற்பொருண்மை மயக்கம்நீக்கம்**	495
9.1.	அறிமுகம்	495
9.2.	சொற்பொருண்மை மயக்கம்	495
9.3.	சொற்பொருண்மை மயக்கநீக்கத்தின் வகைகள்	496
9.3.1.	சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம்	496
9.2.2.	ஒப்புருசொன்மை சார், பல்பொருண்மைசார் பொருண்மை மயக்கம்	497
9.2.3.	மாற்றப் பொருண்மை மயக்கம்	501
9.3.	சொற்பொருண்மை மயக்கநீக்கம்: ஒரு கணினிசார் அணுகுமுறை	501
9.3.1.	புற அறிவு மூலங்கள்	502
9.3.2.	கணினிசார் பொருண்மை மயக்கநீக்கத்தின் வகைகள்	502
9.3.2.1.	கண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம்	503
9.3.2.2.	கண்காணிக்கப்படாத பொருண்மை மயக்கநீக்கம்	503
9.3.2.3.	அறிவு அடிப்படையிலான பொருண்மை மயக்கநீக்கம்	503

9.3.2.4.	பிற அணுகுமுறைகள்	503
9.4.	தமிழில் சொற்பொருள் மயக்க நீக்கத்திற்கான நெறிமுறைகள்	504
9.4.1.	கொத்தாக்கச் சேர்ந்துவருகை அணுகுமுறை	504
9.4.1.1.	சேர்ந்துவருபவைகளின் முக்கியத்துவம்	507
9.4.1.2.	வேற்றுமைக் குறியீடுகளின் முக்கியத்துவம்	507
9.4.1.3.	கற்றல் கட்டம்	508
9.4.1.4.	பரிசோதனைக் கட்டம்	512
9.4.2.	கருத்துரு வரைபட அகராதி அடிப்படையில் சொற்பொருண்மை மயக்கநீக்கம்	514
9.4.2.1	கருத்துவரு வரைபடம்	514
9.4.2.2.	கருத்துரு வரைபட அகராதியைப் பயன்படுத்திப் பொருண்மை மயக்கநீக்கம் செய்தல்	515
9.4.3.	செம்மொழிப் பனுவல்களில் சொற்பொருண்மை மயக்கநீக்கம்	525
9.5.	முடிவுரை	527
10.	இயந்திர மொழிபெயர்ப்பு	529
10.1.	அறிமுகம்	529
10.2.	கணிப்பொறி வழி மொழிபெயர்ப்பின் தேவை	529
10.3.	கணிப்பொறிவழி மொழிபெயர்பின் வகைகள்	530
10.4.	கணிப்பொறி வழி மொழிபெயர்ப்பின் சில முயற்சிகள்	531
10.5.	கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்	533
10.6.	கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்	534
10.7.	இயந்திர மொழிபெயப்பின் வளர்ச்சி	538
10.7.1.	இந்தியாவில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சி	542
10.7.1.1.	அனுசாரக் இயந்திர உதவியிலான மொழிபெயர்ப்பு ஒழுங்குமுறை	543
10.7.1.2.	சிவ மற்றும் சக்தி ஒழுங்குமுறை	546
10.7.1.3.	ஆங்கிலபாரதி	547
10.7.1.4.	அனுபாரதி	552
10.7.1.5.	ஆங்கிலபாரதி II(2004)	552

10.7.1.6.	மந்ரா இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	553
10.7.1.7.	அனுவாதக் மொழிபெயர்ப்பு ஒழுங்குமுறை	554
10.7.1.8.	உலகளாவிய வலைப்பின்னல் மொழி அடிப்படையிலான ஆங்கில இந்திய மொழிபெயர்ப்பு	554
10.7.1.9.	மாத்ரா இயந்திரமொழிபெயப்பு ஒழுங்குமுறை	555
10.7.1.10.	ஆங்கில-கன்னடா இயந்திர மொழிபெயப்பு ஒழுங்குமுறை	555
10.7.1.11.	இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் ஒப்பீடு	556
10.7.2.	தமிழ்சார் இயந்திர மொழிபெயர்ப்புகள்	559
10.7.2.1.	தமிழ் உருஷ்யன் மொழிபெயர்ப்புத் திட்டம்	559
10.7.2.2.	உலக வலைப்பின்னல் மொழி –தமிழுக்கான இடைமொழி இயந்திர மொழிபெயர்ப்பு	561
10.7.2.3.	ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிபெயர்க்கும் திட்டம்	562
10.7.2.4	எயுகேபிசி நிறுவனத்தின் தமிச்-இந்தி மொழிபெயர்ப்புத் திட்டம்	564
10.7.2.5.	இயந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயப்புத் திட்டத்தில் தமிழ் தழுவிய இயந்திர மொழிபெயர்ப்பு	564
10.7.2.8.	கூகுள் மொழிபெயர்ப்பி	574
10.8.	சுருக்க உரை	574
11.	**இயல் 11 : பிற மொழிக் கருவிகள் உருவாக்குதல்**	**576**
11.1.	முன்னுரை	
11.2.	ஒலிபெயர்ப்பி உருவாக்குதல்	576
11.3	எழுத்துப் பெயர்ப்பி உருவாக்குதல்	576
11.4.	எழுத்துணரி உருவாக்குதல்	576
11.5.	சொற்களை அகரவரிசைப் படுத்துதல்	577
11.6.	சொல்லாய்வி/சொல்லாளர் உருவாக்குதல்	577
11.7.	தமிழில் கணிப்பொறி வழி மொழி கற்றல், கற்பித்தல் கருவி உருவாக்கல்	577
11.7.1.	கற்றல், கற்பித்தலில் கணிப்பொறி	579

11.7.2.	கணிப்பொறி வழி கற்றல்-கற்பித்தலின் சிறப்பியல்புகள்	580
11.7.3.	கணினி உதவியுடன் கற்றல் கற்பித்தல் கருவி உருவாக்குவதற்கான கணிப்பொறி மொழிகள்	581
11.7.4.	கணினி உதவியுடன் கற்றல் கற்பித்தல் உருவாக்குவதற்கான பயனீட்டு மென்பொருள்கள்	582
11.7.5.	கணினி உதவியுடன் கற்றல்-கற்பித்தல் கருவியில் நகரும் படங்கள்	583
11.7.6.	கணினி உதவியுடன் கற்றல் கற்பித்தல் கருவி உருவாக்கத்தில் மனித நுண்ணறிவின் பங்கு	583
11.7.7.	கணிப்பொறி வழி மொழி கற்றல், கற்பித்தலின் எல்லைகள்	584
11.7.8.	கணிப்பொறி வழி தமிழ் கற்றல்-கற்பித்தலுக்கான மென்பொருள்கள்	584
11.7.9.	கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் இணையத்தின் பங்கு	585
11.7.10.	மொழி கற்றல் மென்பொருள் உருவாக்கத்தில் பிற துறையினரின் பங்கு	586
11.8.	சுருக்க உரை	587
12	இயல் 12: தமிழின் தொழிநுட்ப வளர்ச்சி	588
12.1.	அறிமுகம்	588
12.2.	சவால்களும் இடையூறுகளும்	589
12.3.	வளங்கள் மற்றும் கருவிகள்	592
12.3.1	மொழிவளங்கள்: வளங்கள், தரவு மற்றும் அறிவுத்தளம்	592
12.3.1.1	உரைத் தரவுத்தொகுதிகள்	592
12.3.1.2.	பேச்சுத் தரவுத்தொகுதிகள்	594
12.3.1.3.	இணைத் தரவுத்தொகுதிகள்	594
12.3.1.4.	சொல்சார் மூலவளங்கள்	595
12.3.1.5.	இலக்கணங்கள்	596
12.3.2.	மொழித் தொழில் நுட்பம்: கருவிகள், தொழில் நுட்பங்கள் பயன்பாடுகள்	597
12.3.2.1.	பேச்சு அறிதல்	597
12.3.2.2.	பேச்சு உருவாக்கம்	598
12.3.2.3.	இலக்கணப் பகுப்பாய்வு	599
12.3.2.3.1.	டோகன்களாகப் பிரித்தல்	600

12.3.2.3.2.	பகுதி ஆய்தல்	601
12.3.2.3.3.	சொல்லன் ஆய்தல்	601
12.3.2.3.4.	உருபனியல் பகுப்பாய்வு	602
12.3.2.3.5.	உருபனியல் உருவாக்கம்	604
12.3.2.3.6.	சொல்வகைப்பாடு அடையாளப்படுத்தல்	605
12.3.2.4.	தொடரியல் பகுப்பாய்வு	606
12.3.2.4.1	கூறாய்வு	607
12.3.2.4.2.	ஆழமற்ற பகுப்பாய்வு	608
12.3.2.4.3.	எச்சத்தொடர் எல்லையை அடையாளம் காணல்	609
12.3.2.5.	ஒலிவழி எழுத்துணர்தல்	610
12.3.2.6.	பொருண்மையியல் பகுப்பாய்வு	611
12.3.2.6.1.	சொற்பொருண்மை மயக்கநீக்கம்	611
12.3.2.6.2.	கேள்விப்பதில் ஒழுங்கமைப்பு	612
12.3.2.6.3.	உறவு பிரித்தெடுத்தல்	613
12.3.2.6.4.	உணர்வுப் பகுப்பாய்வு	614
12.3.2.6.5.	தானியங்கி உரைச்சுருக்கம்	614
12.3.2.6.6.	ஒத்ததைக் குறிப்பிடும் தீர்மானம்	616
12.3.2.6.7.	பெயரிடப்பட்ட இருப்புப்பொருள் கண்டுபிடிப்பு	616
12.3.2.6.8	பிற கணினிப் பொருண்மையியல் ஆய்வுகள்	618
12.3.2.7.	உரை உருவக்கம்	618
12.3.2.8.	இயந்திர மொழிபெயர்ப்பு	619
12.3.2.8.1.	தமிழுக்கு அனுசாரகா	620
12.3.2.8.2.	அனுவாதக்-ஆங்கிலத்திலிருந்து இந்திய மொழிமொழிபெயர்ப்பு செய்யும் ஒரு வலுவான ஒழுங்குமுறை	620
12.3.2.8.3.	இந்திய மொழியிலிருந்து இந்திய மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு	621
12.3.2.8.4.	இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் அமிர்தா விஷ்வ வித்யபீடத்தின் பங்களிப்புகள்	621

12.3.2.8.5.	இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் AUKBCRCயின் பங்களிப்பு	622
12.3.2.8.6.	இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் தமிழ்ப் பல்கலைக் கழகத்தின் பங்களிப்பு	622
12.3.2.8.7.	உத்தமத்தின் பங்களிப்பு	622
12.4.	மதிப்பீடு	623
12.5.	எதிர்கால ஆய்வு மற்றும்வளர்ச்சிக்கான செயல்பாடுகள்	626
	துணைநூல்கள்	628

சுருக்கக்குறியீடுகள்	**விரிவுகள்**
பெதொ	பெயர்தொடர்
விதொ	வினைத் தொடர்
அஅ	அடைகொளி அடை
வேஉதொ	வேற்றுமை உருபுத்தொடர்
பிஉதொ	பின்னுருபுத்தொடர்
வேஉ	வேற்றுமை உருபு
பிஉ	பின்னுருபு
பெஅ	பெயரடை
விஅ	வினையடை
விஅதொ	வினையடைத் தொடர்
பெஅதொ	பெயரடைத்தொடர்

இயல் 1

அறிமுகம்

1.1. நூலின் நோக்கம்

"கணினி மொழியியல்" என்ற தலைப்பில் அமைந்த இந்நூல் கணினி மொழியியல் மற்றும் இயற்கை மொழி ஆய்வின் அடிப்படையான தகவல்களைக் கணினி ஒலியனியல், கணினி உருபனியல், கணினித் தொடரியல், கணினிப் பொருண்மையியல் என்ற தலைப்புகளில் தருவதுடன் இவ்வாய்வினால் ஏற்பட்ட மொழித் தொழிநுட்ப முன்னேற்றங்கள் குறித்தும் விவரிக்கின்றது. தமிழ்த் தொடர்பான கணினி மொழியியல் மற்றும் இயற்கை மொழி ஆய்வு குறித்து இதுவரை வெளிவந்துள்ள ஆய்வேடுகள், கட்டுரைகள், நூல்கள் மற்றும் இணைய தளங்களில் உள்ள தகவல்கள் இவற்றின் அடிப்படையில் செய்திகளைத் திரட்டித் தமிழ் மொழியின் தொழில்நுட்ப வளர்ச்சி பற்றி விளக்கம் தர முயற்சி எடுக்கப்பட்டள்ளது.

1.2. முன்னோடியான நூல்கள்

இந்நூலின் முன்னோடியாகத் தமிழில் கணினியியல் குறித்து வெளிவந்துள்ள நூல்களாக சுப்பையா பிள்ளை (2000) எழுதிய "இயற்கை மொழி ஆய்வு" என்ற நூலையும் பாஸ்கரன் (2004) எழுதிய "தமிழில் கணிப்பொறியியல், கணிப்பொறியில் தமிழ்" என்பனவறைக் குறிப்பிடலாம். சுந்தரம் (2015) எழுதிய கணினித்தமிழும் ஒரு முன்னோடியான நூலாகும்.

வாசு அரங்கநாதனின் "தமிழ்மொழியியலுக்கான கணினிசார் அணுகுமுறைகள்" (Computational Approaches To Tamil Linguistics) மிகவும் வரவேற்கத்தகுந்த பங்களிப்பாகும் (Ranganathan 2016b). கணினி ஒலியியல், கணினி உருபனியயல், கணினி தொடரியல் மற்றும் கணினி பொருண்மையியல் ஆகியவற்றில் தமிழின் மொழியியல் கொள்கைகளைக் கணக்கிடக்கூடிய மற்றும் மொழியியல் கண்ணோட்டத்தில் பொருத்தமான கணினியியல்சார் வழிமுறைகளை வரையறுக்கக்கூடிய நூலாசிரியரால் வடிவமைக்கப்பட்ட ஒரு விரிவான இயற்கை மொழி புரிதல் (Natural Language Understanding (NLU) முறையை இந்தப் படைப்பு முன்வைக்கிறது. இயற்கைமொழி ஆய்வு மற்றும் கணினிகளால் இயற்கை மொழிகளைப் புரிந்துகொள்வது ஆகியவற்றில் நூலாசிரியர் ஒரு பல்துறைகளின் ஒழுங்குமுறையில் கவனம் செலுத்துகிறார். கணினிகளைப் பயன்படுத்தி புதிய வழிகளில் மொழியை எவ்வாறு வடிவமைக்க முடியும் மற்றும் தானியக்கப்படுத்த முடியும் என்பதை விளக்குவதற்கு சமீபத்திய கணினிசார் கோட்பாடுகளும் மொழியியல் கோட்பாடுகளும் மிகவும் முறையான மற்றும் விரிவான முறையில் பயன்படுத்தப்பட்டுள்ளன.

இராசேந்திரன் தாம் தமிழில் எழுதிய இயற்கைமொழி ஆய்வு, கணினிமொழியியல், விரிதரவுமொழியியல், தமிழில் ஒலியியக்கவியலும் உரையிலிருந்து பேச்சும், தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வி, தமிழில் எழுத்துப்பிழைத் திருத்தியும் இலக்கணப்பிழைத் திருத்தியும், தமிழ்ச்சொல்வலை, தமிழில் இயந்திர மொழிபெயர்ப்பும் அதற்கான கருவிகள் உருவாக்கமும் என்ற தமிழ் இயற்கைமொழி ஆய்வு மற்றும் கணினி மொழிமொழியியல்சார் நூல்களை academia.edu மற்றும் Reseach Gate என்ற இணைய தளங்களி பதிவேற்றம் செய்துள்ளார்

1.3. நூலில் அமைப்பு

இந்நூல் பதிமூன்று இயல்களாகப் பகுக்கப்பட்டுள்ளது. முதல் இயல் அறிமுகம் ஆகும். இரண்டாவது இயலில் இயற்கை மொழி ஆய்வு பற்றிய சுருக்கமும் இயற்கை மொழி ஆய்வுக்கு அடிப்படையாக அமையும் சீரான வெளிப்பாடும் தானியங்கிகளும் குறித்தும் விளக்கம் தரப்பட்டுள்ளன. மூன்றாவது இயல் கணினி ஒலியனியல் ஆய்வாகும். இவ்வியலில் ஒலியியக்கவியல் ஆய்வு, உரையிலிருந்து பேச்சாய்வு, பேச்சிலிருந்து உரை ஆய்வு என்பன பற்றியும் தமிழில் இவை குறித்து மேற்கொள்ளப்பட்ட ஆய்வுகள் குறித்தும் விளக்கப்பட்டுள்ளன. நான்காவது இயல் கணினி உருபனியல் ஆய்வாகும். உருபனியல் பகுப்பாய்வு, உருபனியல் உருவாக்கம் என்பன குறித்தும் தமிழில் இவை குறித்து மேற்கொள்ளப்பட்ட ஆய்வுகள் குறித்தும் இவ்வியலில் விளக்கப்பட்டுள்ளன. ஐந்தாவது இயல் கணினித் தொடரியல் ஆய்வாகும். இதில் வாக்கியங்களைப் பகுப்பாய்வது குறித்தும், பலவிதமான பகுப்பாய்வுகள் குறித்தும், தமிழில் வாக்கியங்களைப் பகுப்பாய்வது குறித்து மேற்கொள்ளப்பட்ட ஆய்வுகள் குறித்தும் விளக்கம் தரப்பட்டுள்ளது. ஆறாவது இயல் கணினி பொருண்மையியல் ஆய்வாகும். ஏழாவது இயல் எழுத்துப் பிழை திருத்தமும் இலக்கணப் பிழை திருத்தம் பற்றியதாகும். இவ்வியலில் எழுத்துப் பிழை திருத்தம், இலக்கணப் பிழை திருத்தம் குறித்த அடிப்படை தகவல்கள் தரப்பட்டுள்ளன. மேலும் தமிழில் மேற்கொள்ளப்பட்ட எழுத்துப் பிழை திருத்த முயற்சிகள், இலக்கணப் பிழை திருத்த முயற்சிகள் பற்றியும் விளக்கப்பட்டுள்ளன. எட்டாவது இயல் சொல் மூலங்களை உருவாக்குவது பற்றி விளக்குகின்றது. இதில் சொல் மூலங்களான இயந்திரம் படிக்கவியலும் அகராதி உருவாக்கம், கணினி சொற்களஞ்சிய உருவாக்கம், சொல்வலை உருவாக்கம் என்பன குறித்து விளக்கம் தரப்பட்டுள்ளது. ஒன்பதாவது இயல் சொற்பொருள் மயக்கம் பற்றியதாகும். இதில் சொற்பொருள் மயக்கம் பற்றியும், சொற்பொருள் மயக்க வகைகள் பற்றியும், சொற்பொருள் மயக்கநீக்க வகைகள் பற்றியும் தமிழில் மேற்கொள்ளப்பட்ட சொற்பொருள் மயக்க நீக்க முயற்சிகள் பற்றியும் விளக்கப்பட்டுள்ளன. பத்தாவது இயல் இயந்திர மொழிபெயர்ப்பு குறித்து விளக்குகின்றது. இதில் வேறுபட்ட இயந்திர மொழிபெயர்ப்பு வகைகள் பற்றியும், இந்தியாவில் மேற்கொள்ளப்பட்ட இயந்திர மொழிபெயர்ப்பு முயற்சிகள் பற்றியும் தமிழ் தொடர்பான இயந்திர மொழிபெயர்ப்புகள் பற்றியும் விளக்கப்பட்டுள்ளன.

1.4. நூலின் பயன்பாடு

இது ஒரு தொகுப்பு நூல் என்றாலும் இத்தகைய ஆய்வு அடுத்த கட்ட ஆய்வுகளுக்குப் பயனுள்ளதாய் அமையும். தமிழில் இயற்கை மொழி குறித்த பல ஆய்வுகள் மேற்கொள்ளப்பட்டாலும் வளந்து வரும் இவ்வாய்வுக்களத்தில் ஏற்பட்டுள்ள முன்னேற்றம் குறித்து மாணவர் சமுதாயம் அறியவேண்டும் அவர்கள் இதன் அடிப்படையில் அடுத்த கட்ட உயர்நிலை ஆய்வுகளை மேற்கொள்ள வேண்டும் என்ற ஆர்வத்தைப் தூண்டும் விதமாக இந்நூல் அமைந்துள்ளது. தமிழில் வெளிவந்துள்ள இயற்கை மொழி ஆய்வுகளின் நிறைகுறைகளை ஆய்ந்து மதிப்பீடு செய்யும் ஒரு ஆய்வு மிகவும் தேவைப்பட்டாலும் அத்தகைய முயற்சி இங்கு எடுக்கப்படவில்லை.

1.5. கணிப்பொறி வழி மொழி ஆய்வின் தேவைகளும் சிக்கல்களும்

கணிப்பொறியில் ஆங்கிலத்தைத் தவிர பிற மொழிகளையும் கையாளும் சூழ்நிலை வளர்ந்துள்ளது. குறிப்பாகக் கணிப்பொறியைத் தமிழில் கையாளுதல் கடந்த சில ஆண்டுகளுக்கு முன் தொடங்கப் பெற்று வளர்ச்சி அடைந்து வருகிறது. பொதுவான பயன்பாடுகளுக்குத் தமிழைக் கையாளுதல் கணிப்பொறியில் சமூக மேம்பாட்டை எய்த உதவுகிறது. இது தவிர மொழி ஆய்வுப் பணிகளுக்காகவும் கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன.

1.6. கணிப்பொறியில் மொழிகளைக் கையாளுதல்

கணிப்பொறியில் மொழிகள் இரு விதமாகக் கையாளப் படுகின்றன:

1. அன்றாடப் பணிகளுக்காகக் கணிப்பொறியில் மொழியைக் கையாளுதல்
2. மொழி ஆய்வுக்காகக் கணிப்பொறியைக் கையாளுதல்

முதல் வகைப் பயன்பாட்டில் கணிப்பொறியில் மொழிகள் கையாளுவது கீழ்கண்ட பணிகளுக்கு உதவுகின்றன:

1. தமிழ்ச் சொல்லாய்விகளை உருவாக்குவதன் மூலம் கணிப்பொறியை ஒரு உயர்ரக மின்மத் தட்டச்சு போன்று தமிழுக்குப் பயன்படுத்துதல்
2. டி.டி.பி (Desk Top Printing (DTP)) அதாவது மேஜைமேல் அச்சிடுதல் போன்றவைகளுக்கான மென்பொருள்களைத் தமிழுக்கு உருவாக்கி அச்சுப் பணிகளுக்குக் கணிப்பொறியைப் பயன்படுத்தல்
3. நிறுவனங்களில் பணியாளர்கள் பட்டியல்கள், ஊதியப்பட்டியல்கள் மற்றும் வரவு செலவு அறிக்கைகளைத் தமிழில் தயாரித்தல்
4. வணிக நிறுவனங்களில் பொருள் வாங்கும் மற்றும் விற்கும் பட்டியல்களைக் கணிப்பொறி வழி தமிழில் தயாரித்தல்
5. பலவித செய்தி தொடர்புக்குத் தமிழைக் கையாளுதல்
6. மொழிபெயர்ப்புக் கருவியாகவும் தகவல் பரிமாற்றக்கருவியாகவும் தகவல் பிரித்தெடுத்தல் தகவல் பெறுதல் கருவியாகவும் பயன்படுத்தல்

மேற்கண்ட பணிகளுக்காகக் கணிப்பொறியைக் கையயாளுவது தொடங்கப்பட்டு விரிவடைந்துள்ளது. இத்தகைய வசதிகள் பொதுமக்களுக்குப் பயன்தரத்தக்க வகையில்

அமைந்துள்ளன. எனவே இத்தகைய மென்பொருள்களை வடிவமைப்பதற்கு வணிக ரீதியில் செயல்படும் நிறுவனங்கள் முன் வருகின்றன. அரசும் ஆதரவு தருகிறது. இத்தகைய பணிகளுக்கான மென்பொருள் வடிவமைப்பதற்குத் தற்போது கணிப்பொறிகளில் உள்ள வசதிகள் போதுமானதாகும். மேலும் சாதாரண மொழி அறிவே வடிவமைப்பாளருக்குப் போதுமானதாகும். ஆனால் கணிப்பொறியை மொழி ஆய்விற்குப் பயன்படுத்தும்போது மொழியில் புலமையும் செறிவும் தேவைப்படுகின்றன. இத்தகைய பணிகளைக் கணிப்பொறி வல்லுநர்களும் மொழயியலாளர்களும் இலக்கியவாதிகளும் இணைந்து செயல்பட்டு வடிவமைக்கப்பட வேண்டும். கணிப்பொறியில் தமிழில் உள்ளீடு செய்வதற்கும் வசதிகள் உள்ளன. அதுபோல் வெளியீட்டைப் பெறுவதற்கு உயர்ரக அச்சுப்பொறிகள் உள்ளன. ஆனால் மொழி ஆய்வுப்பணிகளைச் செய்வதற்குப் போதுமான மென்பொருள்கள் இல்லை. இதற்கான முயற்சிகள் தொடங்கப்பட்டு மேம்படுத்தப்பட்டுள்ளன.

1.7. மொழி ஆய்விற்குத் தேவைப்படும் கருவிகள்

மொழியாய்விற்குத் தேவைப்படும் கருவிகளாகப் பின்வருவனவற்றைப் பட்டியலிடலாம்: அகராதிகள், சொற்களஞ்சியங்கள், சொல்வலை, சொல்வங்கி, தகவல்தளம், கிளைப்பட வங்கி, பெருந்தரவு. ஒரு பனுவலின் அல்லது உரையின் சொல்லடைவைக் கணிப்பொறியில் தயாரிக்கலாம். சொல்லடைவில் சொற்கள் சொல்வடிவம் என்ற நிலையில் உள்ளன. இதிலிருந்து அகராதிச் சொற்களைத் தயாரிப்பதற்கு உருபனியல் பகுப்பாய்வு தேவைப்படுகிறது. இப்பகுப்பாய்வினைச் செய்வதற்குச் செயல்முறைகள் கணிதக் கூறுகளாக வடிவமைக்கப்பட வேண்டும். இது எளிய பணியல்ல. இருப்பினும் இப்பணி மொழியாய்வின் தொடக்கப் பணியாக அமைகிறது. தமிழில் அமைந்துள்ள செய்திகளைக் கணப்பொறியில் உள்ளீடு செய்வதற்கு விசைப்பலகை பயன்படுத்தப்படுகிறது. தமிழ் இலக்கிங்களையும் பிற படைப்புகளையும் கணிப்பொறியில் தட்டச்சு செய்து உள்ளீடு செய்ய நிறைய பொருட் செலவும் கால விரையமும் ஆகும். எனவே அச்சு வடிவில் அமைந்த தமிழ்ச் செய்திகளையும் கையெழுத்து செய்திகளையும் கணிப்பொறி படித்து சொல் வடிவங்களாக மாற்ற ஒளிவருடி (Optical Character Recognizer (OCR)) என்ற தொழில் நுட்பம் பயனபடுத்தப்பட்டு ஒளி வருடிக்கான மென்பொருள்கள் உருவாக்கப்பட்டு வருகின்றன. இத்தொழில் நுட்பம் வெற்றியை எட்டிவிட்டது. இக்கருவியியைப் பயன்படுத்தி தமிழ்ச் செய்திகளை, உரைகளை, பனுவல்களை உள்ளீடு செய்வது மிக எளிய பணியாக அமையும். கணிப்பொறிக்குத் தரவேண்டிய கட்டளைகள் மனித மொழியில் தருவதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. ஒரு மனிதன் மற்றொருவரிடம் பேசுவது போன்றும் உரையாடுவது போன்றும் கணிப்பொறியுடன் உரையாடுவதற்கு மென்பொருள் உருவாக்கப்பட்டு வருகின்றன. இப்பணிகளெல்லாம் மொழியியலாளர்களின் மற்றும் கணிப்பொறியாளர்களின் கூட்டு முயற்சியால் நடைபெறுகின்றன. தானியங்கி மொழி பெயர்ப்பிற்கு (automatic translation) ஒரு மொழியின் இலக்கண கூறுகள் அனைத்தையும் கணிப்பொறிக்குத் தர வேண்டி உள்ளது. அதாவது கணிப்பொறிக்கு மூலமொழி மற்றும் இலக்குமொழி இரண்டிலும் தேவையான அறிவினை ஊட்டவேண்டும். இப்பணியும் கணிப்பொறி தயாரிப்பாளரின் கூட்டுமுயற்சியிலே நடைபெறுகின்றன.

1.8. மொழியாய்வில் மொழியியலாளர்களுக்குத் தேவைப்படும் திறமைகள்

மொழி ஆய்வில் மொழியியலாளர்களுக்குப் பின்வரும் திறமைகள் தேவைப்படுகின்றன:

- கணிப்பொறியைப் பற்றிய அடிப்படை அறிவு
- மொழியின் இலக்கணத்தையும் மொழி அலகுகளையும் கணினிப் படுத்துவதில் திறன்
- ஒரு திட்டத்தில் எப்பணிகளைக் கணிப்பொறி வழி செய்ய இயலும், எவ்வாறு செய்ய இயலும் என்பது குறித்த அறிவு
- ஒன்றிணைந்து பணிபுரியும் கணிப்பொறியாளரிடம் மொழி சிக்கல் பற்றி எடுத்துரைக்கும் திறன்
- கலந்துரையாடி முடிவுகள் எடுக்கும் திறன்
- சிக்கலுக்கான தீர்வினைப் படிப்படியாகப் பல்வேறு நிலைகளில் செயல்படுத்துவதற்கான திறன்
- கணிப்பொறியை மொழி ஆய்வுக்குப் பயன்படுத்தும் போது தோன்றும் சிக்கல்களைச் சகமொழியாளர்களிடம் கலந்துரையாடி அதற்குத் தீர்வு காணும் திறன்

1.9. மொழியாய்வில் கணிப்பொறியாளர்க்குத் தேவைப்படும் திறமைகள்

மொழியாய்வில் கணிப்பொறியாளர்களுக்குப் பின்வரும் திறமைகள் தேவைப்படுகின்றன:

- மொழியிலும் மொழி இலக்கணத்திலும் அடிப்படை அறிவு பெற்றிருத்தல்
- மொழியாய்வுப் பணியைப் பல்வேறு சிறு சிறு பணிகளாகப் பிரித்துக் கணிப்பொறி வழி செயல்படுத்துவதற்கான மென்பொருள்களை உருவாக்கும் திறன்
- மொழியாய்வுப் பணிகளுக்கு உதவும் கணிப்பொறி வழியமைப்புகளில் தேர்ச்சி பெற்றிருத்தல்
- மொழியாளருக்குத் தேவைப்படும் மாதிரி வழியமைப்புகளை உருவாக்கி மொழியாளருக்குச் செயல் விளக்கம் தந்து அவர்களின் ஒத்துழைப்பைப் பெறுதல்
- சிக்கல்களுக்கான தீர்வுகளை மொழியாளர்களிடம் இருந்து உடனுக்குடன் எதிர்பாராமல் பொறுமையுடன் காத்திருத்தல்.
- மொழியாளர்களுக்குக் கணிதப்படுத்துவதில் பயிற்சியின்மையால் உடனுக்குடன் தீர்வினைத் தர இயலாது. இந்நிலையில் அவர்களுடன் ஒத்துழைத்து தீர்வு பெற முயற்சித்தல்
- மொழியாய்வுப் பணிகளுக்கான அவ்வப்பொழுது வெளிவரும் மென்பொருள் மற்றும் கணிப்பொறி வசதிகளைத் தெரிந்து கொள்ளுதல்

மேற்கூறியவாறு மொழியியல் ஆய்வுப் பணிகளுக்கு மொழியாளரும் கணிப்பொறியாளரும் சேர்ந்து ஒத்துழைக்கவேண்டும். கணிப்பொறி வழி மொழியாய்வுப் பணியினை மேற்கொள்வதற்கு தேவைப்படும் மென்பொருள்களை உருவாக்க கணிப்பொறியாளர்கள், மொழியாளர்கள், இலக்கியவாதிகள் ஆகியோரின் கூட்டு முயற்சி தேவையாகும்.

1.10. இயற்கை மொழி ஆய்வுகான முன்னேற்பாடுகள்

இந்திய மொழிகள் அனைத்தும் கணிப்பொறி அறிவியலில் உரிய இடத்தைப் பெறுவதற்காக நடுவண் அரசும் மாநில அரசும் பல்கலைக் கழகங்களும் ஆராய்ச்சி நிறுவனங்களும் தனியார் நிறுவனங்களும் பல்வேறு முயற்சிகளில் ஈடுபட்டுவருகின்றன. நடுவண் அரசின் தகவல் தொழில்நுட்ப அமைச்சரகமானது இதற்கென்றே ஒரு தனித் திட்டத்தை – இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சி (Technology Development of Indian Languages (TDIL)) என்ற திட்டத்தை உருவாக்கிச் செயல்பட்டுவருகிறது. இதற்கென்று நாடெங்கும் 13 பல்கலைக்கழகங்கள் மற்றும் ஆய்வு நிறுவனங்களைத் தேர்ந்தெடுத்து, பல கோடி ரூபாய் நிதி உதவி அளித்துச் செயல்படுத்தியது. இத்திட்டத்தின் கீழ் பல இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சிக்கான பல்வேறு ஆய்வுகள் நடைபெற்று வருகின்றன. கடந்த ஆண்டுகளில் தமிழகத்திலும் மாநில அரசானது தமிழ்மொழியின் தொழில்நுட்ப வளர்ச்சிக்காகப் பல்வேறு திட்டங்களைச் செயல்படுத்தியுள்ளது. தமிழ் மென்பொருள் வளர்ச்சிக்கான பல்வேறு திட்டங்களை அறிவித்துச் செயல்படுத்தி வருகிறது. இதன் அடிப்படையில்தான் தமிழ் இணையப் பல்கலைக்கழகமும் நிறுவப்பட்டது. இது பின்னர் இணையக் கல்விக்கழகமாக உருவெடுத்தது. தமிழ் இணைய மாநாடுகள் நடைபெற்று வருகின்றன. தமிழ் மென்பொருள் உருவாக்கத்திற்காக மானிய உதவியைப் பலர் பெற்று செயல்படுத்தியுள்ளனர். சென்னைப் பல்கலைக்கழக மொழியியல் ஆய்வுப் பிரிவு, அண்ணாப்பல்கலைக்கழகக் கணிப்பொறியியல் துறை, சென்னை குரோம்பேட்டையில் அமைந்துள்ள AU–KBC ஆய்வுமையம், எஸ்.எஸ்.என். தொழில்நுட்பக்கல்லூரி, சென்னை, தமிழ் இணைய கல்விக்கழகம், சென்னை, அண்ணாமலைப் பல்கலைக்கழக மொழியியல் துறை, தஞ்சைத் தமிழ்ப்பல்கலைக்கழக மொழியியல் துறை, கோவை பாரதியார் பல்கலைக்கழக மொழியியல் துறை போன்ற பல்வேறு பல்கலைக்கழகத் துறைகளில் தமிழ் மொழியின் தொழில்நுட்ப வளர்ச்சிக்கான பல்வேறு ஆய்வுகள் நிகழ்த்தப்பட்டு வருகின்றன. தமிழில் மென்பொருள் தயாரிப்பாளர்கள் இணைந்து, கணித் தமிழ்ச்சங்கம் என்ற ஒரு அமைப்பைத் தொடங்கி பல்வேறு திட்டங்களை மேற்கொண்டு வருகின்றனர்.

1.11. மொழித்தொழில் நுட்பம்

தொழில்நுட்பம் என்பது மனித வாழ்க்கையின் நடைமுறை நோக்கங்களுக்கு அறிவியல் அறிவைப் பயன்படுத்துதல் ஆகும். இது உயிரித்தொழில்நுட்பம் (bio-technology), வேதியியல் தொழில்நுட்பம் (chemical technology), தோல் தொழில்நுட்பம் (leather technology), தகவல் தொழில்நுட்பம் (information technology) எனப் பலவகைப் படும். அந்த வரிசையில் தற்போது மொழித்தொழில்நுட்பம் (language technology) என்பதும் சேர்ந்துள்ளது. கணினித்துறையில் மனித மொழிகளின் பல்வேறு செயல்பாடுகளுக்காக மென்பொருள் மற்றும் வன்பொருள்களை உருவாக்கும் தொழில்நுட்பமே மொழித்தொழில்நுட்பம் என அழைக்கப்படுகிறது. எடுத்துக்காட்டாக, கணினியில் தமிழைத் தட்டச்சு செய்து, அதைச் சொல் திருத்தி, இலக்கணத் திருத்தி போன்றவற்றின் மூலம் செம்மைசெய்யும் சொல்லாளர் மென்பொருள் (word processor)

மொழித்தொழில்நுட்பத்தின் ஒரு விளைபொருளாகும். அது போன்றே நாம் கணினியில் தட்டச்சுச் செய்வதைப் பேச்சாக மாற்றிக் கொடுக்கும் மென்பொருள் (text-to-speech synthesizer), தட்டச்சு செய்வதற்குப் பதிலாக நாம் கணினியில் பேசினால் அதை எழுத்து வடிவில் தட்டச்சு செய்து கொடுக்கும் மென்பொருள் (automatic speech recognizer), இணைய தளத்தில் நமக்குத் தேவையானவற்றைத் தேடி எடுத்துக் கொடுக்கும் மென்பொருள் (search engine), இவையெல்லாம் மொழித் தொழில்நுட்பத்தின் விளைபொருட்களேயாகும். இவை போன்ற இயற்கை மொழிகளுக்கான மென்பொருள் மற்றும் வன்பொருட்களை உருவாக்கும் மொழித் தொழில்நுட்பத் துறையானது பல்துறை சார்ந்த ஒன்றாகும். கணினித் துறை, தகவல் தொடர்புத்துறை, உளவியல் துறை, மொழியியல் துறை ஆகிய நான்கு துறைகளைச் சேர்ந்த வல்லுநர்களும் இணைந்து செயல்படும் ஒரு துறை இந்தத் மொழித்தொழில்நுட்பத் துறையாகும்.

மேற்கூறிய ஒவ்வொரு துறையினருக்கும் மொழித் தொழில்நுட்பத்தில் ஈடுபடுவதற்கான நோக்கங்கள் வேறுபட்டிருக்கலாம். கணிப்பொறித்துறை வல்லுநர்கள் (computer scientists) கணிப்பொறியைச் செயற்கை அறிவுத் திறனுள்ள ஒரு கருவியாக (artificial intelligent machine) மாற்றும் நோக்கத்தில் மொழித் தொழில்நுட்பத்தில் ஈடுபடலாம். ஏனென்றால் மனிதனின் அறிவுத் திறனுக்கு அடிப்படையாக அவனது இயற்கை மொழி அமைவதால், கணிப்பொறிக்குச் செயற்கை அறிவுத் திறனைக் கொடுக்க முயற்சிக்கும் கணிப்பொறித் துறை வல்லுநர்கள் முதலில் கணிப்பொறிக்கு இயற்கை மொழிகளைப் புரிந்து கொள்ளும் மற்றும் பயன்படுத்தும் திறனைக் கொடுக்க முயலுகின்றனர். அதற்காக அவர்கள் இயற்கை மொழிகளின் பண்பு மற்றும் அமைப்பு என்ன, நமது மனித மூளையானது எவ்வாறு இயற்கை மொழிகளைக் கற்றுக் கொள்கிறது, பயன்படுத்துகிறது போன்றவற்றைக் கண்டறிவதற்காக மொழித் தொழில்நுட்பம் மற்றும் மொழியியல் ஆய்வுகளில் ஈடுபடுகின்றனர்.

தகவல்தொடர்புத்துறை வல்லுநர்கள் (communication engineers) நாம் ஒரு முனையில் இயற்கை மொழிகள் வழியாகப் பேசும் போது அல்லது எழுதும்போது, அதை மின்னணுச்சாதனங்கள் எவ்வாறு எவ்விதச் சொற் சேதமும் தொடர்ச் சேதமும் கருத்துச் சேதமும் இன்றிக் கொண்டுசெல்கிறது அல்லது கொண்டுசெல்லாம் என்பதை ஆய்வதற்காக மொழித் தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

உளவியல் நிபுணர்களோ (psychologists) நாம் மொழி வாயிலாக மற்றவர்களோடு கலந்துரையாடும்போது என்னென்ன உத்திகளைக் கையாளுகின்றோம். இந்த உத்திகளைக் கணிப்பொறி வாயிலாக இயற்கை மொழிகளைப் பயன்படுத்தும் போது எவ்வாறு பயன்படுத்தலாம் என்பதைக் கண்டறிவதற்காகத் மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

மொழியியல் வல்லுநர்கள் (linguists) மொழிகளின் அமைப்பு என்ன, பண்பு என்ன, எவ்வாறு பேச்சொலிகளோ எழுத்துக்களோ கருத்து அல்லது பொருளை வெளிப்படுத்துகிறது என்பவற்றையெல்லாம் கண்டறிந்து, பின்னர் இதனடிப்படையில் கணிப்பொறி வாயிலாக

இயற்கை மொழிப் பயன்பாட்டை மேற்கொள்வதற்காக மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர். அடிப்படையில் மொழித் தொழில்நுட்பத்தின் ஆய்வுப் பொருள் நமது இயற்கை மொழிகளேயாகும். எனவே மொழி அல்லது மொழியியல் ஆய்வின் ஒரு பிரிவே மொழித் தொழில்நுட்பமாகும். எனவே தான் மொழியியலாளர்கள் கணினி மொழியியல் (Computational Linguistics) என்ற ஒரு மொழியியல் ஆய்வுப் பிரிவை உருவாக்கி வளர்த்து வருகின்றனர். அதன் ஒரு உட்பிரிவே மொழித் தொழில்நுட்பமாகும். மொழித்தொழில்நுட்ப ஆய்வில் அல்லது வளர்ச்சியில் மொழி ஆய்வாளர்களுக்கு மிக முக்கியப் பெரும் பங்கு உண்டு. உலக அளவில் கணினி மொழியியல் மற்றும் அதன் உட்பிரிவான மொழித் தொழில்நுட்ப வளர்ச்சியின் பயனாகப் பல்வேறு மொழி மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஜெர்மனியில் ஆச்சிரியப்படத்தக்க ஒரு மொழித் தொழில் நுட்ப வளர்ச்சி அண்மைக் காலத்தில் ஏற்பட்டுள்ளது. அங்கு ஆங்கிலம் மட்டுமே தெரிந்த ஒருவர் சென்றால், அவர் தனது செல்போனில் ஆங்கிலத்திலேயே பேசலாம். மறுமுனையில் ஆங்கிலம் தெரியாத ஒரு ஜெர்மானியர் அதை ஜெர்மானிய மொழியில் கேட்கலாம். ஒரு ஜப்பானியர் தனது ஜப்பானிய மொழியில் கேட்கலாம். அதுபோன்று ஒரு ஜெர்மானியர் அல்லது ஜப்பானியர் தங்களது மொழியில் பேசினால், மறுமுனையில் ஆங்கிலம் மட்டும் தெரிந்தவர் அதை ஆங்கிலத்தில் கேட்கலாம். அதாவது உடனடிக் கணினி மொழிபெயர்ப்பு என்ற ஒரு மொழித் தொழில்நுட்ப வளர்ச்சி அங்கு ஏற்பட்டுள்ளது. கனடா நாட்டில் வானிலையைக் கண்டறியும் கருவிகள் தரும் வானிலை விவரங்களை ஆங்கிலத்தில் தானாக தொகுத்துத் தரும் ஒரு மொழித் தொழில்நுட்பம் உருவாக்கப்பட்டு, தற்போது செயல்படுத்தப்பட்டு வருகிறது. அதுபோன்று விமானம் மற்றும் ரயில் பயணத்திற்கான பதிவுச் சீட்டுகளைப் பதிவு செய்யப்போகின்றவர்களிடம் மனித உதவியாளர்களுக்குப் பதிலாகக் கணிப்பொறியே உரையாடி அவர்களுக்குத் தேவையான பதிவுச் சீட்டைப் பதிவு செய்துதரும் மொழித் தொழில்நுட்ப வளர்ச்சி ஐரோப்பிய நாடுகளில் ஏற்பட்டுள்ளது.

1.12. தமிழ்மொழித் தொழில்நுட்ப வளர்ச்சியின் இன்றைய நிலை

தமிழ் மொழித் தொழில்நுட்பம் நாம் எண்ணிப் பார்க்காத அளவுக்கு வளர்ச்சி அடைந்துள்ளது. கணிப்பொறியில் தமிழைப் பார்க்கவே இயலாத ஒரு கால கட்டம் இருந்தது. கணிப்பொறியில் ஆங்கிலம் போன்ற மொழிகளில் மட்டுமே தட்டச்சு செய்யவோ அல்லது பிற கணிப்பொறிப் பயன்பாடுகளை மேற்கொள்ளவோ முடிந்தது. பின்னர் தமிழ் எழுத்துருக்களை (Tamil fonts) கணிப்பொறியில் கொண்டு வரும் முயற்சி வெற்றிபெற்றது. அதன் பயனாகக் கணிப்பொறியில் தமிழைத் தட்டச்சு செய்யும் வளர்ச்சி ஏற்பட்டது. அதிலும் பல சிக்கல்கள் நீடித்தன. தமிழுக்குப் பல வேறுப்பட்ட கணினித் தட்டச்சுப் பலகைகள் அறிமுகப்படுத்தப்பட்டன. மேலும் குறிப்பிட்ட ஒரு தமிழ் மென்பொருளில் தட்டச்சு செய்தலை மற்றொரு மென்பொருள் கொண்டு திருத்தவோ மாற்றவோ செய்ய இயலாத நிலை நீடித்தது. பின்னர் தமிழக அரசின் உதவியுடன் நடைப்பெற்ற தமிழ் இணையம் மாநாடு ஓரளவுக்கு இச்சிக்கல்களைத் தீர்த்து வைத்தது. தமிழுக்கென யுனிக்கோட் உருவாக்கப்பட்டு பயன்பாட்டில் உள்ளது. இதுபோன்ற முயற்சிகளினால் கணிப்பொறிக்கும் நமக்கும் இடையில் தமிழ் பயன்படத் தொடங்கியது. தமிழைக்

கணிப்பொறியில் பார்க்கத் தொடங்கினோம். அதாவது பயன்பாட்டாளர் இடையூடகம் (user interface) என்ற அடிப்படையில் தமிழ் பயன்படத் தொடங்கியது. பல கணினி மென்பொருள்களில் தமிழ் எழுத்துக்கள் நடைபோடத் தொடங்கின. இந்த வளர்ச்சியைத் தமிழ் வாயிலாகக் கணிப்பொறியின் பயன்பாடு என்று அழைக்கலாம். தமிழுக்கு என்று சொல்லாளர்கள் (word processor) பல உருவாக்கப்பட்டு பயன்பாட்டில் உள்ளன. பேச்சறிவான் (speech recognizer), எழுத்து-பேச்சு மாற்றி (text-to-speech transferer), இணையத் தேடுவான் (search engine) போன்ற பலமொழிசார் மென்பொருள்கள் தமிழுக்கு உருவாக்கப்பட்டு வருகின்றன. சொற்பிழை அல்லது இலக்கணப் பிழை ஏற்பட்டால் தானாக அவற்றைத் திருத்தித்தரும் சொல்லாளர்கள் உருவாக்கப்படு வருகின்றன. இதுபோன்ற தமிழ் மென்பொருள்களைத் (software) தமிழுக்கான மென்பொருட்கள் என்று அழைக்கலாம். இவ்வகை மென்பொருட்கள் உருவாக்கப்படுவதில் இருந்த தடைகள் தற்போது நேர் செய்யப்பட்டுள்ளன. துறைசார் மென்பொருட்கள் உருவாக்கப்பட்டு வருகின்றன. மருத்துவத்துறை மென்பொருட்கள் உருவாக்கப்பட வேண்டும் என்றால், முதலில் குறிப்பிட்ட மருத்துவத்துறையில் அறிவு உடையோர் அத்துறையில் ஆராய்ச்சி மேற்கொண்டு, மருத்துவ உண்மைகளைத் தெளிவுப்படுத்தும் போதுதான் அவற்றை அடிப்படையாகக் கொண்டு, மருத்துவர்களின் உதவியோடு கணிப்பொறித்துறை வல்லுநர்கள் மருத்துவ மென்பொருட்களை உருவாக்க இயலும். அதுபோல ஒரு மொழிக்கான மென்பொருளை உருவாக்கவேண்டும் என்றால், மொழி அறிஞர்கள் அக்குறிப்பிட்ட மொழியின் அமைப்பு, பண்பு, செயல்பாடு போன்றவற்றை ஆராய்ந்து வெளிப்படுத்தும்போது தான், அவற்றை அடிப்படையாகக் கொண்டு கணிப்பொறித்துறை வல்லுநர்கள் அம்மொழிக்கு மென்பொருட்களை உருவாக்க இயலும். அவ்வாறு இல்லாமல் அம்மொழி பேசும் கணிப்பொறித்துறை வல்லுநர்கள் தாங்களே அம்மென்பொருட்களை உருவாக்க முயன்றால் அது அவ்வளவு வெற்றி பெறாது. மொழி வல்லுனர்களின் துணையின்றி கணிப்பொறி வல்லுனர்கள் மொழிசார் மென்பொருள்களைச் செம்மையாக உருவாக்கவியலாது என்பது தான் உண்மை.

தமிழுக்கு ஒரு நல்ல சொல்திருத்தி (spell checker) உருவாக்கப்படவேண்டும் என்றால், தமிழ்ச் சொற்களின் பண்பு மற்றும் அமைப்பைப் பற்றி உண்மைகளைத் தெளிவாகத் தமிழ்மொழி அறிஞர்கள் முன்வைக்கும்போதுதான், அதை அடிப்படையாகக் கொண்டு, கணிப்பொறித்துறை வல்லுநர்கள் சொல்திருத்தி மென்பொருளைத் தமிழுக்கு உருவாக்க முடியும். தமிழில் பெயர்ச்சொற்கள் எந்தெந்த விகுதிகளை எடுக்கும், எந்த வரிசையில் எடுக்கும், எவ்வாறு அவற்றை இணைத்து ஒரு முழுச் சொல்லை உருவாக்குவது, எவ்வாறு தமிழ்ச் சொற்களைப் பகுபத உறுப்புகளாகப் பிரிப்பது போன்ற உண்மைகளை எல்லாம் தெள்ளத்தெளிவாக, அறிவியல் பூர்வமாகத் தமிழ் மொழி அறிஞர்கள் எடுத்துச்சொன்னால், கணிப்பொறித்துறை வல்லுநர்கள் அவற்றை அடிப்படையாகக் கொண்டு சொல்திருத்தி மென்பொருட்களை உருவாக்க இயலும்.

இந்த இடத்தில் ஓர் உண்மையைத் தெளிவுப்படுத்த வேண்டியது அவசியம். மனித மூளை மொழியைப் புரிந்துகொள்ளும் முறை வேறு, கணிப்பொறி புரிந்துகொள்ளும்முறை வேறு. அதற்குக் காரணம், மனித மூளை ஒரு உயிர் வேதியியல் பொருள்; ஆனால் கணிப்பொறியில் உள்ள சிப் (chip) ஒரு மின்னணுப்பொருள். எனவே எந்த ஒரு அறிவையும் மனித மூளைக்குக் கொடுக்கும் முறை வேறு, மின்னணுச் சிப்புக்குக் கொடுக்கும் முறை வேறு. இவ்வேறுபாட்டை மனதில் கொண்டு மொழி அறிஞர்கள் தாங்கள் கண்டறிந்த உண்மைகளைக் கணிப்பொறிக்குப் புரியும் வகையில் மாற்றி அமைத்துகொடுத்தால் தான் கணிப்பொறி வல்லுநர்களின் வேலை எளிதாகும். இதுபோன்ற இலக்கணங்களைக் கணினி இலக்கணம் (computational grammar) என்று அழைப்பார்கள். தமிழில் இத்தகைய முயற்சி மேற்கொள்ளப்பட்டு வெற்றி அடைந்துள்ளன. தற்போது பல்வேறு நோக்கில் பல்வேறு கணினி இலக்கணங்கள் உருவாக்கப்பட்டுள்ளன. இதன் காரணமாக நல்ல சொல்லாளர்கள், பேச்சறிவான்கள், எழுத்து-பேச்சு மாற்றிகள், இணையத் தேடுவான்கள், இயந்திர மொழிபெயர்ப்பிகள் போன்ற மென்பொருட்கள் உருவாக்கப்பட்டு செயல்படுத்தப்பட்டு வருகின்றன. இவ்வாறு கணிப்பொறி நோக்கில் மொழியாய்வு மேற்கொள்ளப்படுவதுதான் கணினி மொழியியல் என்று அழைக்கப்படுகிறது. அதன் உட்பிரிவே மொழித் தொழில்நுட்பம் ஆகும். தமிழ்க் கணினி மொழியியல் வளர்ச்சி அடைந்து தமிழுக்கான பல மென்பொருட்கள் உருவாக்கப்பட்டு வருகின்றன. இதுவே கணிப்பொறியில் உயர்நிலைத் தமிழ்மொழிப் பயன்பாடாகும். தமிழ் எழுத்துருக்களை (fonts) உருவாக்கி, கணிப்பொறியில் பிற மென்பொருட்களோடு உறவாடும் நிலையிலிருந்து (user interface) அடுத்தக் கட்ட நிலைக்கு நாம் சென்றுகொண்டிருக்கின்றோம். ஆங்கிலம் போன்று உயர்நிலை மென்பொருட்கள் உருவாக்கப்பட்டு வருகின்றன. கணினி மொழியியலும் அதன் உட்பிரிவான மொழித் தொழில்நுட்பமும் இன்று உலகளவில் நன்கு வளர்ச்சி அடைந்துள்ளன. இதற்கென்று தனித்துறைகள் பல்வேறு பல்கலைக்கழகங்களில் உருவாக்கப்பட்டுச் செயல்பட்டுவருகின்றன. தமிழ் மொழியிலும் கணினிமொழியியல் கோட்பாடுகளின் அடிப்படையில் பல ஆய்வுகள் மேற்கொள்ளப்பட்டு வருகின்றன. தமிழுக்குச் சொற்பகுப்பான் (morphological parser), சொல் உருவாக்கி (word generator), தொடர் பகுப்பான் (syntactic parser), தொடர் உருவாக்கி (sentence generator) போன்ற அடிப்படை மொழித் தொழில்நுட்ப மென்பொருள் கருவிகள் உருவாக்கப்படுவதும் பௌதிக ஒலியியல் அடிப்படையிலான தமிழ்ப்பேச்சொலி ஆய்வு, பொருண்மையியல் (semantics) அடிப்படையிலான ஆய்வு, விரிதரவு மொழியியல் ஆய்வு (corpus linguistics) போன்றவை மேற்கொள்ளப்படுவதும் கணினித் தமிழ் வளர்ச்சிக்கு இன்றியமையாதன. இவை இல்லாமல் தமிழுக்கு உயர்நிலையில் மென்பொருட்கள் உருவாக்க இயலாது. கணிப்பொறிமயமாக உள்ள இன்றைய உலகில், உயர்நிலைத் தமிழ்மென்பொருட்கள் தமிழுக்கு உருவாக்கப்படாமல் உலக அளவில் மற்ற மொழிகளுக்கு இணையாகத் தமிழ் அடுத்த கட்ட

வளர்ச்சியை எட்டவியலாது. அச்சு இயந்திரம் தோன்றியபோது, தமிழ் அதில் இடம் பெற்ற காரணத்தால்தான் இன்று தமிழ் எழுத்துலகில் நீடித்து வளர்ந்து வந்துள்ளது என்பதை நாம் மறந்துவிடலாகாது. அதுபோன்று கணிப்பொறியில் உரிய இடத்தைத் தமிழ் பெற்றால்தான் அடுத்த கட்ட வளர்ச்சிக்குத் தமிழ் செல்லும் என்பதை நாம் நினைவில்கொள்ள வேண்டும். தமிழுக்கு எழுத்துருவை உருவாக்கிக் கணிப்பொறியில் பார்ப்பதோடு நிறுத்திக்கொள்ளாமல், பல்வேறு உயர்நிலைத் தமிழ்மென்பொருள் உருவாக்கத்தில் நாம் கவனம் செலுத்தவேண்டும். எனவே, தமிழ் ஆசிரியர்களும் ஆய்வாளர்களும் மாணவர்களும் கணினிமொழியியல் மற்றும் மொழித் தொழில்நுட்பம் அடிப்படையில் தமிழ் ஆய்வை மேற்கொள்ளவேண்டும். உயர்நிலைத் தமிழ் மென்பொருட்கள் உருவாக்கத்திற்குத் தேவையான தமிழுக்கான அடிப்படை மொழித் தொழில்நுட்ப மென்பொருள் கருவிகள் உருவாக்கப்படவேண்டும்.

1.13. உலகளவில் கணினி மொழியியலின் தொடக்கம் வளர்ச்சியும்

உலகளவில் கணினி மொழியியலின் தொடக்கம் வளர்ச்சியும் குறித்து மிகச்சுருக்கமான தகவல்கள் கீழே தரப்பட்டுள்ளன (Ruth Camburn. 2013).

- 1940கள்: கணினி மொழியியல் பிறந்தது; தொடக்க காலத் திட்டங்கள் இயந்திர மொழிபெயர்ப்பில் கவனம் செலுத்தின.
- 1954: ரஷ்ய-ஆங்கில இயந்திர மொழிபெயர்ப்பு முறையின் முதல் பொதுச் செயல் விளக்கம் (அது தோல்வியடைந்தாலும் கூட)
- 1962: பிரபல கணினி அறிவியலாரும் தானியங்கி மொழிச் செயலாக்க ஆலோசனைக் குழுவின் (ஆல்பாக்) உறுப்பினருமான டேவிட் ஹேஸ் “கணினி மொழியியல்” என்ற வார்த்தையை உருவாக்கினார்
- 1964: எம்ஐடி பேராசிரியர் ஜோசப் வீசன்பாம் மனித உரையாடலைப் பிரதிபலிக்கும் முதல் உரையாடலான எலிசாவை உருவாக்கினார்
- 1980கள்: கணினி மொழியியல் பற்றிய ஆராய்ச்சியின் முதல் நடைமுறை வணிகமயமாக்கல் மற்றும் இயந்திர மொழிபெயர்ப்புக்குத் திரும்புதல்
- 1990கள்: பெரிய நிறுவனங்களின் சிறந்த வளர்ச்சி மற்றும் இயந்திர கற்றலின் பயன்பாடு
- 2000கள்: இயற்கை மொழி ஆய்வுக் கருவிகள் மற்றும் இடைமுகங்களின் திடீர் எழுச்சி.

இயல் 2

கணிப்பொறி மூலம் மொழி ஆய்வு

2.1 இயற்கை மொழி ஆய்வு

கணிப்பொறி ஒரு மொழியாய்வுக் கருவி. செய்திகளை 1,0 என்ற எண்களால் ஆய்வது தான் கணிப்பொறி. மனித மூளை 0,1 என்ற அடிப்படை எண்களை வைத்துக் கொண்டு எந்திர மொழி மூலம் சிக்கலான கணிப்பொறியை உருவாக்கியது. எழுத்துக்களும் எண்களும் தனித்தன்மையான குறியீடுகளும் (special symbols) 0,1 என்பவைகளால் தான் உருப்படுத்தம் செய்யப்படுகின்றன. உயர்நிலை மொழிகள், பொருள்கோளி (interpreters), ஒன்று கூட்டிகள் (assemblers) என்பன மொழியியலால் பயன்பெற்றன. இப்பொழுது மொழியியலார் கணினியை அவர்களின் கடினமான செயல்பாடுகளைச் செய்யும் உண்மையான கருவியாகப் பயன்படுத்தும் காலம் வந்துவிட்டது. மொழியியல் சார் கணிப்பொறி ஆய்வைக் கணினி மொழியியல் ஆய்வு அல்லது இயற்கை மொழி ஆய்வு என்பர். இது பேச்சைத் தெரிந்துகொள்ளுதல், பேச்சை உருவாக்குதல், பனுவல்களை/உரைகளைப் பேசாக்குதல், பேச்சுக்களைப் உரை ஆக்குதல், இயந்திர மொழிபெயர்ப்பு, விரிதரவுகளை ஆய்தல், உரைகளை ஆய்தல், மொழி அலகுகளைக் பகுப்பாய்வு செய்தல் என்பனவற்றை உள்ளடக்கும். மொழியாய்வு மொழியைப் புரிந்துக் கொள்ளுதல் என்பதுடன் பின்வருவனவற்றையும் உள்ளடக்கும்: உரை அல்லது பேச்சுச் செய்திகளை பொருள் கொண்டுப் புரிந்துக் கொள்ளுதல். மொழிபெயர்த்தல், உரைகள் அல்லது பேச்சுச் செய்திகளை உருவாக்குதல்.

2.2 இயற்கை மொழியும் செயற்கைமொழியும்

மனிதர்கள் இயற்கையாகப் பயன்படுத்தும் எந்த மொழியும் இயற்கை மொழியாகும். செயற்கையாக மனிதனால் உருவாக்கப்பட்ட மொழிகள் இயற்கை மொழியில் இருந்து வேறுபட்டவை. எடுத்துக்காட்டாக, கணிப்பொறியில் பயன்படுத்தும் வழிமுறைவரைவு மொழிகளான (programming languages) போர்ட்ரான் (Fortran), பேசிக் (Basic), சி (C), புரோலாக் (Prolog), லிஸ்ப் (Lisp), பெர்ல் (Perl) என்பவைகளெல்லாம் செயற்கை மொழிகளாகும். இயற்கை மொழியாய்வு (Natural Language Processing (NLP)) கணிப்பொறியைக் கொண்டு மொழியை ஆய்வு செய்யும் எல்லா நடவடிக்கைகளையும் உள்ளடக்கும். பேச்சையும் (speech) எழுத்து மொழியையும் ஆயும் கணினி நுட்பங்கள் இதில் அடங்கும். பேச்சு உருவாக்கம் (speech synthesis) இயற்கை ஒலியான பேச்சை உருவாக்குவது நுட்ப அடிப்படையில் சிக்கலானதாகும். என்ன பேசப்படுகிறது என்பதைப் புரிந்து கொள்ள சரியான இசையோட்டம் போன்றவை தேவை.

2.3 மொழி ஆய்வும் கணிப்பொறியும்

இன்றைய அறிவியல் உலகத்தின் நாயகன் கணிப்பொறி. இருபதாம் நூற்றாண்டில் அறிமுகப்படுத்தப்பட்ட பல்வேறு பயன்பாட்டுச் சாதனங்களைப் பின் தள்ளிவிட்டு நமக்கு இன்றியமையாத பங்காளியாய் மாறியுள்ளது கணிப்பொறி. வெறும் கணிக்கும் பொருளாக இருந்த கணிப்பொறி தரவுகள் சேமித்து வைக்கும் கருவியாக, தகவல் தொடர்பு கருவியாக, தரவுகளை தானே ஆய்ந்து நமக்கு முடிவுகளை வெளிப்படுத்தும் கருவியாக மற்றும் பலவழிகளில் பயன்படுகிறது. மொழி செயல்பாட்டிற்கும் கணிப்பொறியின் பங்கு நம்மை வியப்படையச் செய்யும் விதத்தில் வளர்ந்துள்ளது. தற்காலக் கணிப்பொறியின் நினைவகம் மிகப்பெரிய அளவில் உள்ளது. துணை நினைவகங்களும் மிகப்பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கானச் சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில் (நினைவகம், பதிவுத்தட்டு, பதிவுநாடா போன்றவற்றில்) சேகரித்து வைத்துத் தேவையானபோது பயன்படுத்திக் கொள்ள இயலும். மொழிக்கான தரவுகள் பலவகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவோ வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். இவைகள் உள்ளடங்கிய ஒரு பனுவலாகவும் (text) அமையலாம். வெவ்வேறு வகைப்பட்ட உரைகளிலிருந்து தேர்ந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு பெருந்தரவு/விரிதரவு/தரவுத்தொகுதி (corpus) என்று அழைக்கப்பெறும். தமிழுக்கான பெருந்தரவைச் சேகரித்து வைப்பது மிக இன்றியமையாதப் பணியாகும். பல இலக்கியக் கூறுகளை உள்ளடக்கிய உரைகள், பல நடையியல் கூறுகளை உள்ளடக்கிய உரைகள், வட்டார வழக்கு, கிளை மொழி வழக்கு, சமுதாயமொழி வழக்கு, பேச்சு வழக்கு, இலக்கிய வழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய உரைகள் எல்லாம் ஒருசேரப் பெருந்தரவாக அமையும். இத்தகைய பெருந்தரவை இருவகையாகப் பிரிக்கலாம்:

1. எழுத்து வடிவில் அமையும் பெருந்தரவு (written corpus)
2. பேச்சு வடிவில் அமையும் பெருந்தரவு (speech corpus)

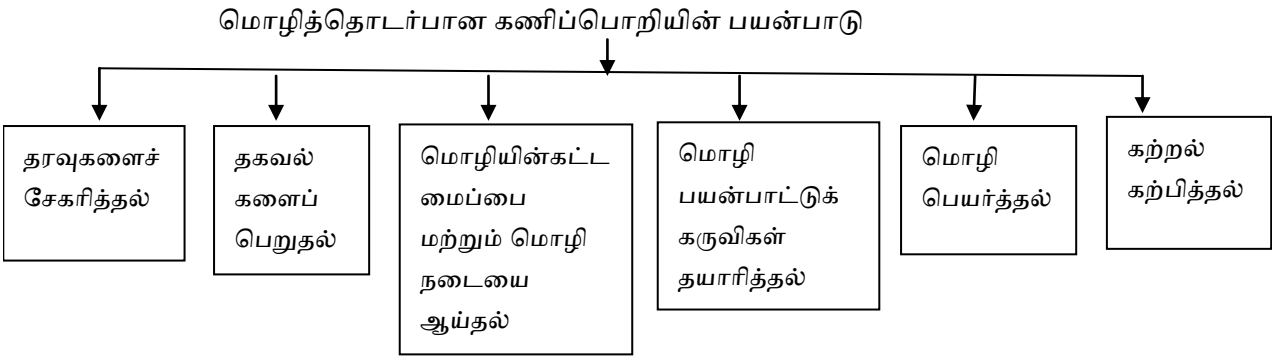
இத்தகைய பெருந்தரவுகள் மொழியாய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவிப் புரிகின்றன. தரவுகள் கணிப்பொறியால் ஆயப்பட்டு தகவல்களாகத் தரப்படும். மொழித்தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வன்தட்டு, குறுந்தட்டு, பதிவுநாடா போன்றவற்றில்) சேமித்து வைத்துத் தேவைப்படும்போது பயன்படுத்திக் கொள்ளலாம். பல்வேறு விதமான தகவல்கள் இணையத்தளத்தில் பெருமளவில் நிரம்பிக் கிடக்கின்றன. இத்தகவல்களைப் பெறுவதில் முக்கியமான அம்சம் என்னவென்றால் இயல்மொழியைப் பயன்படுத்தி தேவையான தகவல்களைப் பெறுதலாகும். இதற்கு தேவை இயல்மொழி இடைமுகப்பு (Natural language Interface). இதனால் இயல்மொழி அடிப்படையிலான குறியீடுகளைக் கொண்டு தேவையான தகவல்களைப் பெற இயலும். தமிழ் மொழியின் கட்டமைப்பைக் கணிப்பொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை, கருத்தாடல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ் மொழியைப் பயன்படுத்தி தகவல்களைப் பெற உதவும்

இயல்மொழி இடைமுகப்புகளை நிறுவ இயலும். இத்தகைய இடைமுகப்பு பல்வேறு வடிவுகளைக் கொண்டிருக்கும். தகவல்தேடும் பொறி (search engine) தகவல்களைப் பெற உதவும் முக்கியக் கருவியாகும். தமிழுக்கு இத்தகைய தகவல்தேடும் பொறி உருவாக்குவதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டு பெருமளவில் வெற்றியும் அடைந்துள்ளது. கேள்வி-பதில் அமைப்புகள், கணிப்பொறியுடன் உரையாடும் அமைப்புகள் போன்றவை இயல்மொழி இடைமுகப்பை அடிப்படையாகக் கொண்டு அமைக்கப்பட்டிருந்தால் தான் பாமரரும் பயன்படுத்தி தகவல்களைப் பெற இயலும். எடுத்துக்காட்டாக, தமிழ் இலக்கிய நூல் ஒன்றை விடுவிரலாக தந்து இயல்மொழி முகப்பு மூலம் தேவையான பதில்களை அல்லது தகவல்களை பெறும்படியாக கணிப்பொறியின் பயன்பாடு அமையலாம். விவசாயிகளுக்கான தரவு மையங்களிலிருந்து இயல்மொழியின் வழி வேண்டிய தகவல்களைப் பெறுதலும் இத்தகைய பயன்பாடாகும். மொழி உரைகளைப் படிநிலை (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) அமைப்பில் வரும் மொழி அலகுகளாக பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்துக் கொள்ள கணிப்பொறியைப் பயன்படுத்தலாம். பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக் கொள்வோம்: *பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்*. *பள்ளி மாணவன் ரகுமான்* என்பது பெயர்த்தொடர் என்றும், *பலாப்பழத்தைக் கவனமாகப் பறித்தான்* என்பது வினைத்தொடர் என்றும், மேலும் *பள்ளி மாணவன்*, *மரம்*, *பலாப்பழம்*, *ரகுமான்* என்பன பெயர்கள் என்றும், *பறி* என்பது வினை என்றும், *பெரிய* என்பது பெயரடை என்றும், *கவனமாக* என்பது வினையடை என்றும், *இலிருந்து* பின்னுருபு என்றும், *ஐ* என்பது வேற்றுமையுருபு என்றும், *-த்-* என்பது இறந்தகால இடைநிலை என்றும், *ஆன்* என்பது படர்க்கை ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிபொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களைப் பகுப்பாய்வு செய்து தந்துவிடும். ஒரு காலகட்டத்தில் மொழியலார் கள ஆய்வு செய்து தகவலாளிகளின் உதவியால் ஆயப்படாத மொழிகளின் தரவுகளைச் சேகரித்து அதன் மீது சில மொழியியல் கொள்கைகளையும் கோட்பாடுகளையும் பயன்படுத்தி சில கருதுகோள்களின் அடிப்படையிலும் உற்றுநோக்கல் மூலமும் அம்மொழிகளின் ஒலியனியல், உருபனியல், தொடரியல் அமைப்புகளைக் கண்டபிடிக்க முயலுவர். இது பண்புசார் மொழியியல் (qualitative linguistics) எனப்படும். இப்போது கணிப்பொறியின் வரவால் அத்தகைய நிலை மாறி மொழிகளின் விரிதரவுகள் உருவாக்கப்பட்டு அதன்மீது புள்ளியில் நெறிமுறைகளின் அடிப்படையில் மேற்சொன்ன மொழி அமைப்புகள் கண்டுபிடிக்கப்படும். இதை அளவுசார் மொழியியல் (quantitative linguistics) என்பர். தற்போது மொழி ஆய்வுகள் எல்லாம் கணினியைப் பயன்படுத்தி விரிதரவுகள் மீது செய்யப்படும் ஆய்வுகளாக மாறியுள்ளன. மொழி உரையை இடு வரலாகத் (input) தந்து தக்க வழியமைப்புகளும் (programs) கட்டளைகளும் (instructions) தந்தால் மொழியின் கட்டமைப்பு குறித்த தகவல்களை விடுவரலாகப் (output) பெறலாம். கணிப்பொறியைக் கொண்டு மொழி உரைகளை

வாக்கியங்களாகவும் வாக்கியங்களைத் தொடர்களாகவும் தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும் ஒலியன்களை ஒலிக்கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளுக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்காகத் தொடரியல் ஆய்வி (syntactic analyzer), உருபனியல் ஆய்வி (morphological analyzer), ஒலியனியல் ஆய்வி (phonological analyzer), பொருண்மையியல் ஆய்வி (semantic anayazer) என்ற ஆய்விகளைத் தயாரிக்க வேண்டும். தமிழில் இத்தகைய ஆய்விகள் தயாரிக்கப்பட்டுப் பயன்பாட்டில் உள்ளது. இவ்வாறு மொழிக் கூறுகளைக் கணிப்பொறி வழி பகுப்பாய்வது குறித்தும் தமிழில் இதற்கான முயற்சிகள் குறித்தும் இவ்வாய்விகள் அடிப்படையிலான மொழிசார் மென்பொருள்கள் அல்லது மொழிக் கருவிகள் உருவாக்குவது குறித்தும் இந்நூலில் விளக்கப்படும்.

மொழித் தொடர்பான கணிப்பொறியின் பயன்பாட்டை ஆறு முக்கியக் கூறுகளாகப் பகுக்கலாம்.

படம் 1



முதல் இரண்டும் மொழி அல்லாத பிற நெறிகளுக்கும் பொருந்தும். மூன்றாவது பயன்பாடு பின்னர்வரும் மூன்று பயன்பாட்டிற்கும் அடிப்படையாக அமையும்.

2.3.1 தரவுகளை சேகரித்தல்

தற்காலக் கணிப்பொறியின் நினைவகம் மிகப் பெரிய அளவில் உள்ளது. துணை நினைவகங்களும் மிகப் பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கான சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில் (நினைவகம், பதிவுத் தட்டு, பதிவு நாடா போன்றவற்றில்) சேகரிப்பு வைத்துத் தேவையான போது பயன்படுத்திக்கொள்ள இயலும்.

மொழிக்கான தரவுகள் பல வகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவொ வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். ஒரு பனுவலாக (text) அமையலாம். (எந்த ஒரு மொழி உடலும் பனுவலாகும்.) வெவ்வேறு வகைப்பட்ட பனுவல்களில் இருந்து தெரிந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு பெருந்தரவு (corpus) என்று அழைக்கப்பெறும். தமிழுக்கான பெருந்தரவைச் சேகரித்து வைப்பது மிக

இன்றியமையாத பணியாகும். பல இலக்கியக் கூறுகளை உள்ளடக்கிய பனுவல்கள், பல நடையியல் கூறுகளை உள்ளடக்கிய பனுவல்கள், வட்டார வழக்கு, கிளைமொழி வழக்கு, சமுதாய மொழி வழக்கு, பேச்சு வழக்கு, இலக்கிய வழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய பனுவல்கள் எல்லாம் ஒரு சேரப் பெருந்தரவாக அமையும். இத்தகைய பெருந்தரவை இருவகையாகப் பிரிக்கலாம். ஒன்று எழுத்து வடிவில் அமையும் பெருந்தரவு (written corpus) மற்றொன்று பேச்சு வடிவில் அமையும் பெருந்தரவு (speech corpus). இத்தகைய பெருந்தரவுகள் மொழி ஆய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவி புரியும்.

2.3.2 தகவல்களைப் பெறுதல்

தரவுகள் கணிப்பொறியால் ஆயப்பட்டுத் தகவல்களாகத் தரப்படும். மொழித்தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வன் தட்டு, குறுந்தட்டு, பதிவு நாடா போன்றவற்றில்) சேமித்து வைத்துத் தேவைப்படும் போது பயன்படுத்திக்கொள்ளலாம். பல்வேறு விதமான தகவல்கள் இணையதளத்தில் பெருமளவில் நிரம்பிக் கிடக்கின்றன. இத்தகவல்களைப் பெறுவதில் முக்கியமான அம்சம் என்னவென்றால் இயல் மொழியைப் பயன்படுத்தித் தேவையான தகவல்களைப் பெறுதலாகும். இதற்குத் தேவை இயல் மொழி இடைமுகப்பு (Natural Language Interface (NLI). இதனால் இயல் மொழி அடிப்படையிலான குறியீடுகளைக் கொண்டே தேவையான தகவல்களைப் பெற இயலும். தமிழ் மொழியின் கட்டமைப்பைக் கணிப்பொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ்மொழியைப் பயன்படுத்தித் தகவல்களைப் பெற உதவும் இயல்மொழி இடைமுகப்புகளை நிறுவ இயலும். இத்தகைய இடைமுகப்பு பல்வேறு வடிவுகளைக் கொண்டிருக்கும். தகவல் தேடும் பொறி (search engine) தகவல்களைப் பெற உதவும் முக்கிய கருவியாகும். தமிழுக்கு இத்தகைய தகவல் தேடும் பொறி உருவாக்குவதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டும் அவை பெருமளவில் வெற்றியும் அடைந்துள்ளன.

2.3.3 மொழியின் கட்டமைப்பை ஆய்தல்

மொழிப் பனுவல்களை ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை ஆகியப் படிநிலை அமைப்பில் வரும் மொழி அலகுகளாகப் (ஒலியன்கள், உருபன்கள், சொற்கள், தொடர்கள், வாக்கியங்கள், கருத்துரை, தர்க்கக் கூற்று) பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்துகொள்ளக் கணிப்பொறியைப் பயன்படுத்தலாம்; மறுதலையாகச் சிறிய மொழிக் கூறுகளிலிருந்து பெரிய மொழிக் கூறுகளை ஆக்கலாம். முதலாவதைப் மொழிப் பகுப்பாய்வு என்றும் இரண்டாவதை மொழி ஆக்கம் என்றும்

அழைக்கலாம். சிக்கலான மொழி அமைப்பைப் பின்வருமாறு ஆறு வேறுபட்ட ஆய்வுகளாகப் பகுக்கலாம்:

- ஒலியனியல் ஆய்வு
- உருபனியல் ஆய்வு
- தொடரியல் ஆய்வு
- பொருண்மையியல் ஆய்வு
- பயன்வழியியல் ஆய்வு
- கருத்தாடல் ஆய்வு

மொழியை இம்மட்டங்களில் பகுப்பதைக் கணிப்பொறி வழி மொழிப் பகுப்பாய்வு என்பர். மறுதலையை மொழி ஆக்கம் என்பர். இவை கணினி மொழியியலுக்கு அடிப்படை ஆய்வுகளாகும்.

ஒலியனியல் ஆய்வு: ஒலியனியல் ஆய்வு சொற்களை அல்லது உருபன்களை ஒலியன்களாகப் பகுப்பாய்வு செய்யும்; ஒலி அடையாளக் குறிகளிடமிருந்து அல்லது கேட்பு அடையாளக் குறிகளிடமிருந்து ஒலியன்களைப் பிரித்தெடுக்கும்; மேலும் பேச்சைத் தெரிந்து கொள்ளும் மற்றும் ஆக்கம் செய்யும்.

ஒலிகள் ↔ சொற்கள்

/b/+/ɔ:/ +/t / ↔ /bɔ:t/ ‘boat’

க்+ஆ+க்+அ+ம் ↔ காலம்

உருபனியல் ஆய்வு: இது சொற்களை உருபன்பகளாகப் பகுத்தல், உருபன்களிலிருந்து சொற்களைப் பெறல் என்பனவற்றை உள்ளடக்கும்.

உருபன்கள் ↔ சொற்கள்

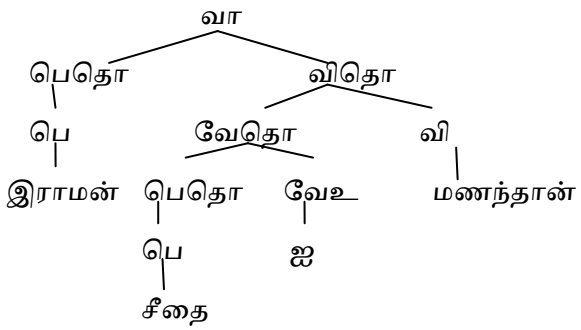
மரம்+கள் ↔ மரங்கள்

தொடரியல் ஆய்வு: இது வாக்கியங்களைத் தொடர்களாகவும் சொற்களாகவும் பிரித்தல், சொற்களின் வரிசை அமைப்பிலிருந்து வாக்கிய அமைப்பைப் பெறல் என்பனவற்றை உள்ளடக்கும்.

சொல்வரிசை ↔ வாக்கிய அமைப்பு

‘இராமன் சீதையை மணந்தான்’ என்ற சொல் வரிசையிலிருந்து பின் வரும் பகுத்தாயப் பட்ட கிளை அமைப்பைப் பெறலாம்.

படம் 2:



வா = வாக்கியம், பெதொ = பெயர்த் தொடர், விதொ = வினைத் தொடர், வேதொ = வேற்றுமைத் தொடர், பெ = பெயர், வி = வினை, வேஉ = வேற்றுமை உருபு.

பொருண்மையியல் ஆய்வு: இது வாக்கிய அமைப்பு, சொற்பொருண்மை இவற்றிலிருந்து வாக்கியப் பொருண்மையைப் பெறுவதை உட்படுத்தும். எடுத்துக்காட்டாக, *இராமன் சீதையை மணந்தான்* என்று பகுத்தாயப்பட்ட கிளை அமைப்பிலிருந்து பின்வரும் தர்க்கக் கூற்றைப் பெறலாம்:

மணந்தான் (*இராமன்*, *சீதை*).

பயன்வழியியல் ஆய்வு: வாக்கியப்பொருள் மற்றும் சூழல் இவற்றிலிருந்து சரியான பொருளைப் பெறுவதை உட்படுத்தும். இது பொது அறிவையும் முன்னர் வந்த கூற்றின் அறிவையும் உள்ளடக்கும். *அவன் சீதையை மணந்தான்* என்ற வாக்கியத்தில் வரும் பதிலீடுப் பெயர் (அவன்) முன்னர் வந்த வாக்கியத்தின் அறிவால் அவன் = ராமன் என்று புரிந்துகொள்ளப்படும்.

பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக்கொள்ளவும்.

பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்.

பள்ளிமாணவன் ரகுமான் என்பது பெயர்த்தொடர் என்றும் *மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்* என்பது வினைத் தொடர் என்றும் மேலும் *பள்ளி, மாணவன், மரம், பலாப்பழம், ரகுமான்* என்பன பெயர்கள் என்றும் *பறி* என்பது வினை என்றும் *பெரிய* என்பது பெயரடை என்றும் *கவனமாக* என்பது வினையடை என்றும் *இலிருந்து*, *ஐ* என்பன வேற்றுமை உருபுகள் என்றும் *த்* என்பது இறந்த கால இடைநிலை என்றும், *ஆன்* என்பது படர்க்கை ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிப்பொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களைப் பகுத்தாய்வு செய்து தந்து விடும்.

மொழிப் பனுவலை இடுவராலாகத் (input) தந்து தக்க வழியமைப்புகளும் (programs) கட்டளைகளும் தந்தால், மொழியின் கட்டமைப்பு குறித்த தகவல்களை விடுவரலாகப் (output) பெறலாம். கணிப்பொறியைக் கொண்டு மொழிப்பனுவல்களை வாக்கியங்களாகவும் வாக்கியங்ளைத் தொடர்களாகவும் தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும்

ஒலியன்களை ஒலிக் கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளுக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்காகத் தொடரியல் பகுப்பாய்வி (syntactic analyzer/parser), உருபனியல் பகுப்பாய்வி (morphological analyzer/parser), ஒலியனியல் பகுபாய்வி (phonological analyzer), பொருண்மையியல் பகுப்பாய்வி (semantic analyzer/parser) என்ற பகுப்பாய்விகளைத் உருவாக்க வேண்டும். தமிழில் இத்தகைய ஆய்விகள் உருவாக்கப்பட்டுப் பயன்பாட்டில் உள்ளன. இதற்கு மறுதலையாகச் செயல்படும் கருவிகளை ஆக்கிகள் எனலாம்; தொடரியல் ஆக்கி, உருபனியல் ஆக்கி என்பன ஆக்கிகளாகும்.

2.3.4 மொழி நடையை ஆய்தல்

மொழியின் கட்டமைப்பைக் கருத்தாடல், வாக்கியம், தொடர் சொல் என்ற அலகுகளின் அடிப்படையில் கணிப்பொறி மூலம் ஆய இயலும் என்று கண்டோம். மொழி அலகுகள் நடையியல் அடிப்படையிலான சிறப்புப் பண்புக் கூறுகளை வெளிப்படுத்தி மொழிநடை வேறுபாட்டை உணர்த்தும். கணிப்பொறியின் உதவியால் வேறுபட்ட நடையில் அமைந்த பனுவல்களை அல்லது மொழி வழக்குகளை ஆய்ந்து ஒப்பீடு செய்து அவற்றின் நடையியல் சிறப்புப் பண்புகளைத் தெரிந்து கொள்ளலாம். இத்தகைய ஆய்வு படைப்பாளியை இனங்கண்டு கொள்ளவும் படைப்பிலக்கியத்தின் காலத்தைக் கணிக்கவும் பயன்படும். கணிப்பொறி உதவியுடன் இத்தகைய ஆய்வை நுணுக்கமாகவும் விரைவாகவும் செய்து முடிக்கலாம். இத்தகைய, கணினி வழி தமிழ் மொழி நடை ஆய்வுகள் பல வெற்றிகரமாகச் செய்து முடிக்கப்பட்டுள்ளன. புள்ளியல் மொழியியல் (Statistical Linguistics), அளவீட்டு மொழியியல் (Quantitative Linguistics) போன்ற பயன்பாட்டு மொழியியல் பாடங்களின் கோட்பாடுகளும் கொள்கைகளும் மொழியியல் நடை ஆய்வுக்கு உறுதுணை புரிகின்றன. மொழிநடையை ஆய உதவும் மென்பொருள்கள் பல கண்டு பிடிக்கப்பட்டு மொழிப் பயன்பாட்டில் உள்ளன.

2.3.5. பேச்சைத் தெரிந்து கொள்ளல்

இது அடிப்படையில் தொடர்ச்சியான ஒலி அலைகளை வேறுபட்ட சொற்களாகப் பகுத்தல், இயற்கைமொழியைப் புரிந்துக் கொள்ளுதல் (Natural Language Understanding), தனித்தனிச் சொற்களிலிருந்து பொருண்மைக்குச் செல்லுதல், இயற்கை மொழியை உருவாக்குதல், (Natural Language Generation), அனுமானிக்க இயலாத உள்ளீடுகளுக்குப் பொருத்தமான பதில்களை உருவாக்குதல் என்பனவற்றை உள்ளடக்கும்.

2.3.6 இயந்திர மொழிபெயர்ப்பு

இது ஒரு இயற்கை மொழியை இன்னொன்றாக மொழிப்பெயர்ப்பு செய்தலைக் குறிப்பிடும். பொதுவாக இயற்கை மொழியாய்வு பேச்சை உட்படுத்தாமல் தான்

பயன்படுத்தப்படுகின்றது. பேச்சு மற்றும் இயற்கை மொழியாய்வு என்பதைப் பேச்சையும் பிற இயற்கை மொழியாய்வுகளையும் உள்ளடக்கப் பயன்படுத்தலாம்.

2.4 சீரான வெளிப்பாடுகளும் தானியங்கிகளும்

கணினியின் வழி மொழி ஆய்வுகள் செய்ய சீரான வெளிப்பாடுகளையும் தானியங்கிகளையும் (Regular expressions and Automata) பற்றி நமக்கு அடிப்படையான அறிவு தேவை.

2.4.1 சீரான வெளிப்பாடுகள் (Regular expressions)

பனுவல் தொடர் வரிசைகளைப் (text strings) பண்பாக்கம் செய்யும் தரமான குறிமானம் (standard notation) சீரான வெளிப்பாடாகும் (regular expressions). சீரான வெளிப்பாடு வலையில் தேடல் போன்ற சூழல்களில் பனுவல் கோர்வைகளை (text strings) குறிப்பிட்டுக் கூறவும் பிற தகவல் மீட்பு (information retrieval) பயன்பாடுகளிலும் பயன்படுத்தப்படுவதோடு சொல் ஆய்விலும், விரிதரவுகளிலிருந்து (corpora) வருகை எண்ணிக்கைகளைக் கணக்கிடல் போன்ற செயல்பாடுகளிலும் முக்கியமான பங்கு வகிக்கின்றது. கணிப்பொறி அறிவியலில் தரப்படுத்தல்களின் ஒரு வெற்றி சீரானவெளிப்பாடுகளாகும்; இது உரைதேடல் கோர்வைகளைக் குறிப்பிட உதவும் ஒரு மொழியாகும்.

சீரான வெளிப்பாடு முதலில் க்லின் (Kleen 1956) என்பவரால் உருவாக்கப்பட்டது. இது கோர்வைகளின் எளிய வகுப்புகளை குறிப்பிட பயன்படுத்தப்படும் ஒறி தனித்தன்மையான மொழியில் உள்ள வாய்பாடாகும் (formula). ஒரு கோர்வை என்பது குறியீடுகளின் வரிசையாகும். பனுவல் அடிப்படையிலான தேடல் நுட்பங்களுக்கு ஒரு கோர்வை என்பது எண்ணெழுத்துக்களின் (alphanumeric characters) (எழுத்துக்கள், எண்கள், இடைவெளிகள், தாவல்கள், நிறுத்தற்குறியீடுகள்) ஒரு வரிசையாகும். இந்த நோக்கத்திற்கு ஒரு இடைவெளி என்பது ஒரு எழுத்தாகும்; நாம் இதை ‘ ‘ என்ற குறியீட்டால் உருப்படுத்தம் செய்யலாம்.

முறையாக ஒரு சீரான வெளிப்பாடு ஒரு குழும கோர்வைகளை பண்பாக்கம் செய்யும் ஒரு இயல்கணித குறிமுறையாகும் (algebraic notation). இவ்வாறு அவற்றை தேடல் கோர்வைகளைக் குறிப்பிடவும் ஒரு மொழியை வாய்ப்பாட்டு வழியில் வரையறுக்கவும் பயன்படுத்த இயலும். கோர்வைகளைப் பண்பாக்கம் செய்ய மூன்று சீரான வெளிப்பாட்டு வினைக்குறிகள் (regular expression operators) போதுமானதாகையால் பெர்ல் மொழியின் வசதியான மற்றும் பொதுவாகப் பயன்படுத்தப்படும் சீரான வெளிப்பாட்டுத் தொடரியலை (regular expression syntax) பயன்படுத்த இயலும். பனுவல் செயற்பாங்கு திட்டவரைவுகள் (text-processing programs) பெரும்பாலான சீரான வெளிப்பாட்டுத் தொடரியலுடன் உடன்படும் ஆகையால் நாம் குறிப்பிடும்

பெரும்பாலானவை UNIX, Microsoft Word, WordPerect ஆகியவற்றின் சீரான வெளிப்பாடுகளுக்கு நீட்சிசெய்யும் (Jurafsky and Martin 2000: 22).

சீரான வெளிபாட்டுத் தேடல் (regular expression search) நாம் தேட விரும்பும் அமைப்பொழுங்கையும் தேடுவதற்கு பனுவல்களின் விரிதரவையும் வேண்டும். சீரான வெளிப்பாட்டு தேடல் செயல்பாடு விரிதரவு முழுவதும் தேடி அந்த அமைப்பொழுங்கு இருக்கும் எல்லா உரைகளையும் எடுத்துத் தரும். வலைத் தேடல் பொறி (Web search engine) போன்ற தகவல் மீட்பு (information retrieval (IR) ஒழுங்குமுறையில் (system) பனுவல்/உரை முழு ஆவணமாகவோ வலைப் பக்கங்களாகவோ இருக்கக்கூடும். ஒரு சொல் ஆய்வியில் (word processor), பனுவல்/உரை தனிச் சொல்லாகவோ ஒரு ஆவணத்தின் வரிகளாகவோ இருக்கக்கூடும். UNIX-இன் greb என்ற கட்டளை இதைத்தான் செய்கின்றது. ஒரு தேடலை ஒரு சீரான வெளிப்பாட்டுடன் பொருந்துபவைகளை எடுத்துத்தரும்படியோ அல்லது முதல் பொருத்தத்தை எடுத்துத் தரும்படியோ வடிவமைக்கலாம்.

2.4.1.1. மிக அடிப்படையான சீரான வெளிப்பாட்டு அமைப்பொழுங்கு

மிக எளிதான சீரான வெளிப்பாட்டு வகை தனி எழுத்துக்களின் ஒரு தொடர்வரிசையாகும். எடுத்துக்காட்டாக, Wood Chuck என்பதைத் தேட நாம் /Wood Chuck/ என்று தட்டச்சு செய்கின்றோம். எனவே /Butter cup/ என்ற சீரான வெளிப்பாடு Buttercup என்ற துணைக் கோர்வையை (substring) உள்ளடக்கும் எந்த கோவையுடனும் பொருத்தம் பெறும். தேடல் கோர்வை (search string) ஒரு எழுத்தையோ (எடுத்துக்காட்டு: / ! /)அல்லது ஒரு தொடர் வரிசையான எழுத்துக்களைக் கொண்டிருக்கலாம் (எடுத்துக்காட்டு: /urgl/). சீரான வெளிப்பாட்டுடன் ஒவ்வொரு பொருத்தத்தின் முதல் நிகழ்வும் பின்வரும் அட்டவணையில் கோடிடப்பட்டு தரப்பட்டுள்ளது (Jurafsky and Martin 2000: 23).

சீரான வெளிப்பாடுகள்	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
/woodchuks/	“Intersting links to woodchucks and lemurs”
/a/	“Mary Ann stoped by Mona’s”
/Claire_ says,/	“Dagmor my gift please“, Claire says,”
/song/	“all our pretty songs”
/!/	“You’ve left the burgular behind again! said Nori

சீரான வெளிப்பாடுகள் எழுத்துகள் சின்ன வடிவ வகையைச் சார்ந்தனவா பெரிய வடிவ வகையைச் சார்ந்தனவா என்ற உணர்வு உள்ளன. எடுத்துக்காட்டக சின்ன பெரிய வடிவ வகையைச் சார்ந்தனவா என்ற உணர்வு உள்ளன. எடுத்துக்காட்டாக சின்ன வடிவ வகையைச் சார்ந்த /s/ /S/ /woodchuk/ / Woodchuk/ (disjunction) (Ww) . சீரான வெளிப்பாடுகள் பொருத்தம் பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டுகள் பின்வரும் அட்டவணையில் தரப்பட்டுள்ளன (Jurafsky and Martin 2000: 24).

சீரான வெளிப்பாடுகள்	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
/(Ww)oodchuk/	woodchucks அல்லது Woodchuck	“woodchuck”
/[abc]/	‘a’, ‘b’, or ‘c’	“Near the pond”
/[123456789]/	ஏதாவது ஒரு எண்	planty of 7 to 5

சீரான வெளிப்பாடான /[1234567890]/ என்பது எந்த ஒரு ஒற்றை எண்ணையும் குறிப்பிடும். எழுத்துக்களை பின்வருவாறு குறிப்பிடுவது நன்றாக இராது.

[ABCDEFGHIJKLMNOPQRSTUVWXYZ]

இங்கு எழுத்துக்களை ‘_‘ என்ற குறியுடன் பயன்படுத்தலாம். எடுத்துக்காட்டாக /[b-g]/என்பது b,c, d,e,f,g, என்ற எதாவது எழுத்தைக் குறிப்பிடும். இது போல /[2-5] / என்ற அமைப்பொழுங்கு 2,3,4,5 என்ற எதாவது ஒற்றை ஸ்தான எண்ணை குறிப்பிடும். சில எடுத்துக்காட்டுகள் பின்வரும் அட்டவணையில் தரப்பட்டுள்ளன (Jurafsky and Martin 2000: 24).

சீரான வெளிப்பாடுகள்	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
/[A-Z]/	பெரிய வடிவ வகை எழுத்து	“We met Ramasamy’
/[a-z]/	சின்ன வடிவ வகை எழுத்து	“my book”
/[0-9]/	ஒரு இலக்க எண்	“Chapter1”

‘^’ (‘காரட் ‘caret’) என்ற குறியீட்டின் துணையுடன் செவ்வக அடைப்புக் குறியை எந்த ஒரு எழுத்து இல்லை என்பதைக் குறிப்பிடவும் பயன்படுத்தலாம். எடுத்துக்காட்டாக /[^a]/ என்பது a - ஐத் தவிர வேறு எந்த எழுத்துடனும் பொருத்தமுறும் ‘^’ என்பது முதல் குறியீடாக இருக்கும் போது தான் இது உண்மையாகும். சில எடுத்துக்காட்டுகள் பின்வரும் அட்டவணையில் தரப்பட்டுள்ளன (Jurafsky and Martin 2000: 24)

சீரான வெளிப்பாடுகள்	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
/[^A-Z]/	பெரிய வடிவ வகை எழுத்து அல்ல	“Online service”
[^Ss]	‘S’- ஓ ‘s’-ஓ அல்ல	“I have a pen.”
[^\.]	முற்றுப்புள்ளி அல்ல	“our book”
[e^]	‘e’ அல்லது ‘^’	“look for ^ now”
a^b	‘a^b’ என்ற அமைப்பொழுங்கு	“look for a^b now”

எவ்வாறு நாம் woodchuck என்பதையும் woodchucks என்பதையும் குறிப்பிட இயலும்? என்பது தற்போதைய பிரச்சினை. இதற்கு நாம் கேள்விக்குறியை /?/ பயன்படுத்தலாம்; கேள்விக்குறி முன்வரும் எழுத்தையோ ஒன்றும் இல்லாமையோ (பூஜியத்தையோ) குறிப்பிடும். சில எடுத்துக்காட்டுகள் பின்வரும் அட்டவணையில் தரப்பட்டுள்ளன (Jurafsky and Martin 2000: 25)

சீரான வெளிப்பாடுகள்	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
Woodchucks	woodchuck அல்லது woodchucks	“woodchuck”
Colou?r	color அல்லது colour	“colour”

எத்தனை தடவை எழுத்து வர வேண்டும் என்பதை எவ்வாறு குறிப்பிடுவது என்பது நமது அடுத்த பிரச்சனை. சில வேளைகளில் சீரான வெளிப்பாடு திரும்பத்திரும்ப வருதலையும் அனுமதிக்க வேண்டும். எடுத்துக்காட்டாக ஆட்டின் மொழியை எடுத்துக் கொண்டால் பின்வரும் கோர்வைகள் அமைந்திருக்கும்.

baa!

baaa!

baaaa!

baaaaa!

baaaaaa!

b என்பதைத் தொடர்ந்து குறைந்தது இரண்டும் ‘a’ வர இறுதியில் ஆச்சரியக்குறி வரும் கோர்வை இம்மொழியில் அமைந்திருக்கும். திரும்பத்திரும்ப வருவதைக் குறிப்பிட நாம் க்ளீன் ஸ்டார் * (Kleene Star *) என்று பொதுவாக அழைக்கப்படும் நட்சத்திரக் குறியீட்டின் (asterisk அல்லது *) அடிப்படையில் அமைந்த ஒரு குழும (set of operators) இயக்கிகளைப் பயன்படுத்தலாம். க்ளீன்

ஸ்டார் “முன்வரும் எழுத்து பூஜியத் தடவையோ பலதடவையோ வருதல்” என்பதை உணர்த்தும். எடுத்துக்காட்டாக /a*/ என்பது ‘a’ பூஜிய தடவையோ பல தடவையோ வரும் கோர்வை” என்று பொருள்படும். இது a என்பதோடும் aaaa என்பதோடும் பொருத்தமுறும். மட்டுமன்றி பூஜிய தடவை a வரும் “Often” என்பதோடும் பொருத்தமுறும். எனவே ஒரு தடவையோ பல தடவையோ a வருவதைக் குறிப்பிடும் சீரான வெளிப்பாடு /aa*/ என்பதாகும். மிகச் சிக்கலான அமைப்பொழுங்கையும் திரும்பக்கூற இயலும். எடுத்துக்காட்டாக /[ab]*/ என்பது “பூஜியம் அல்லது பல தடவை a-க்கள் அல்லது b-க்கள் வரும் என்று பொருள்படும். இது aaaa அல்லது ababab அல்லது bbbb என்ற கோர்வைகளுடன் பொருத்தமுறும். இன்டிஜர் (integer) (ஸ்தான எண்களின் கோர்வை) என்பதன் சீரான வெளிப்பாடு /(0-9) (0-9)*/ என அமையும். சீரான வெளிப்பாட்டில் குறைந்தது ஒரு தடவை என்று குறிப்பிட வேண்டி வரும். க்ளினெட் (Klneenet) என்பதைப் பயன்படுத்தினால் முந்திய எழுத்து ஒரு தடவையோ பல தடவையோ” என்று பொருள்படும். எடுத்துக்காட்டாக /(0-9)+/ என்பது “ஸ்தான எண்களின் கோர்வை என்பதைக் குறிப்பிடும் சாதாரண வழியாகும். இவ்வாறு ஆட்டின் மொழியை இரு வகைகளில் குறிப்பிடலாம். /baaa*!/ அல்லது /baa+!/ மிக முக்கியமான பிரத்தேகமான எழுத்து முற்றுப்புள்ளி (/. /) ஆகும். இது எந்த ஒரு எழுத்துடனும் பொருத்தமுறும் ஏகோபித்த ஒன்றாகும் (wild card). பின்வரும் அட்டவணையில் எடுத்துக்காட்டு தரப்பட்டுள்ளது (Jurafsky and Martin 2000: 26).

சீரான வெளிப்பாடுகள்	பொருத்தம்	பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு
/beg.n/	Any character between beg after and n	begin, beg’n, began, begun

ஏகோபித்த எழுத்து (.) கிளின் ஸ்டாருடன் “ஏதாவது எழுத்துகளின் கோர்வை” என்ற பொருளில் பயன்படுத்தப்படும். எடுத்துக்காட்டாக நாம் *aardvak* என்ற சொல் இரண்டுதடவை வரும் ஒரு வரியைக் கண்டுபிடிக்கவேண்டுமானால் நாம் இதைச் சீரான வெளிப்பாட்டுடன் பயன்படுத்தலாம்: /aardvark.* aardvark/ (Jurafsky and Martin 2000: 26).

நங்கூரங்கள் (anchors) ஒரு குறிப்பிட்ட இடங்களில் நங்கூரங்கள் சீரான வெளிப்பாட்டை ஒரு கோர்வையின் குறிப்பிட இடங்களில் நிறுவும் தனிப்பட்ட எழுத்துகளாகும். காரட்டும் (caret=^) டாலர்குறியும் (dolorsign $) மிகப் பொதுவான நங்கூரங்களாகும். காரட் ^ ஒரு வரியின் ஆரம்பத்துடன் பொருத்தமுறும். எடுத்துக்காட்டாக /^The/ The என்ற சொல்லுடன்

பொருத்தமுறும். காரட்டுக்கு இவ்வாறு மூன்று பயன்பாடுகள் உண்டு (Jurafsky and Martin 2000: 26)..

1. வரியின் தொடக்கத்துடன் பொருத்தமுறும்.
2. செவ்வக அடைப்புக்குறிக்குள் மறுதலிக்கும்.
3. காரட் என்பதைக் குறிப்பிடும்.

டாலர் குறி ($) ஒரு வரியின் இறுதியுடன் பொருத்தமுறும். எடுத்துக்காட்டாக $ என்ற அமைப்பொழுங்கு வரியின் இறுதியில் வரும் இடைவெளியுடன் பொருத்தமுற பயன்படும். /^The dog\.$/ என்பது The dog என்ற தொடர் உள்ள வரியுடன் மட்டுமே பொருத்தமுறும். (.) என்பது முற்றுப்புள்ளி தான் ஏகோபித்த எழுத்தல்ல என்பதைக் குறிப்பிட இடச் சாய்கோடு (Slash) பயன்படுத்தப்படுகின்றது (Jurafsky and Martin 2000: 26).

மேலும் இரண்டு நங்கூரங்கள் உண்டு. \b என்பது சொல் எல்லையுடன் பொருத்தமுறும். \B சொல் எல்லை அல்லாமையுடன் பொருத்தமுறும். எடுத்துக்காட்டாக/\b The \b/என்பது The என்ற சொல்லுடன் பொருத்தமுறும்; Other என்ற சொல்லுடன் பொருத்தமுறாது. நுணுக்கமாக பெர்ல் ஒரு சொல்லை இலக்க ஸ்தான எண்கள் (Digits) அல்லது கீழ்க்குறியீடுகள் (underscores) அல்லது எழுத்துகள் (letters) என்பதன் வரிசை என்று விளக்குகின்றது. இது பெர்ல், 'சி' என்ற வழியமைப்பு மொழிகளில் (programing Language) 'word' என்பது விளக்கப்பட்டுள்ளதை அடிப்படையாகக் கொண்டது. எடுத்துக்காட்டாக /\b99/ என்பது There are 99 bottles என்ற வாக்கியத்திலுள்ள 99 என்ற கோர்வையுடன் தொடர்புறும்; ஆனால் There are 299 bottles என்ற வாக்கியத்திலுள்ள 99 என்ற கோர்வையுடன் தொடர்புறாது; ஆனால் $99 என்பதிலுள்ள 99-உடன் தொடர்புறும். (& ஒரு ஸ்தான எண்ணோ அடிக் குறியீடோ எழுத்தோ அல்ல) (Jurafsky and Martin 2000: 26)..

2.4.1.2. பிரித்தல் (disjunction) குழுமுதல் (grouping) மற்றும் முன்நிலை (Precedence)

பிரித்தலுக்கு பைப் குறியீடு (pipe symbol "|") என்ற பிரித்தல் இயக்கி பயன்படுகின்றது. எடுத்துக்காட்டாக /cat |dog/ என்ற அமைப்பொழுங்கு cat அல்லது dog என்ற கோர்வையுடன் பொருத்தமுறும். பிரித்தல் குறியீடு ஒரு குறிப்பிட்ட அமைப்பொழுங்கில் மட்டும் செயலுற பிறை (வில்) அடைப்புக்குறி இயக்கிகளைப் பயன்படுத்த வேண்டிவரும். எடுத்துக்காட்டாக /gupp (y|ies)/ என்ற அமைப்பொழுங்கு பிரித்தல். y-க்கும் ies -க்கும் இடையே தான் செயல்படும் என்பதைக் குறிப்பிடும். ஒரு இயக்கி வேறு ஒரு இயக்கிக்கு முன்னர் செயல்படும் என்பது சீரான வெளிப்பாட்டின் இயக்கி முன்னிலைப்படுகை படிநிலை அமைப்பு என்பதால் குறிப்பிடப்படும். பின்வரும் அட்டவணை கூடுதல் முன்னிலைப்படுகையிலிருந்து குறைந்த முன்னிலை படுகை

வரை சீரான வெளிப்பாட்டு இயக்கி முன்னிலைப் படுகையின் நிரக்கைத் தரும் (Jurafsky and Martin 2000: 28).

பிறை அடைப்புக் குறிகள்	( )
எண்ணிகள்	* + ? { }
வரிசைகளும் நங்கூரங்களும்	The ^ my end $
பிரிதல்	|

எடுத்துக்காட்டாக எண்ணிகளுக்கு வரிசைகளை விட கூடுதல் முன்னிலைப்படுகை இருப்பதால், /The*/ என்பது 'theeeee' என்பதுடன் பொருத்தமுறும்; The the என்பதுடன் பொருத்தமுறாது.

2.4.1.3.எளிதான எடுத்துக்காட்டுகள் (Jurafsky and Martin 2000: 28).

நாம் ஆங்கிலச் சொல் the என்பதை சீரான வெளிப்பட்டில் எழுதவேண்டுமென்றால் மிக எழுமையான அமைப்பொழுங்கு கீழ்வருவதாகும்:

/the/

ஒரு சிக்கல் என்னவென்றால் ஆங்கிலத்தில் சொல்லானது வாக்கியத் தொடக்கத்தில் பெரிய வடிவில் வரும். இதற்கு வேண்டி நாம் பின்வரும் அமைப்பொழுங்கைப் பயன்படுத்தலாம்.

/[tT]he/

சில வேளைகளில் the என்பது other, theology என்ற சொற்களுக்குள் வரலாம். எனவே the என்பதன் முன்னும் பின்னும் எழுத்துக்கள் வரவில்லை என்பதைக் குறிப்பிட நாம் இருபக்கமும் சொல் எல்லையை பின்வருமாறு குறிப்பிட வேண்டும்:

/ \b [Th] he \b/

/(^a-zA,Z) (tT) he (^a-z A-Z)/

/(^a-zA,Z) (tT) he (^a-z A-Z)/

2.4.1.4.கலவைத்தன்மையான எடுத்துக்காட்டுகள்

நாம் சீரானவெளிப்பாடுகளின் திறன் காரணியின் (power of REs) சிறப்பான எடுத்துக்காட்டைக் காண்போம். விலைகளுக்கான சீரான வெளிப்பட்டை எடுத்துக்கொள்வோம். இங்கு ஒரு சீரான வெளிப்பாடாக ஒரு டாலர் குறியீட்டைத் தொடர்ந்து இலக்கங்களின் கோர்வை வரும். பெர்லில் $ வரியின் இறுதியைக் குறிப்பிடது என்பதைக் குறித்துக்கொள்ளவும்; இதை பின்வருமாறு அறியலாம் (Jurafsky and Martin 2000: 29).

/ $ [0 - 9] + /

இப்போது நாம் டாலர்களின் பின்னங்களைக் கையாளவேண்டும். இதற்கு நாம் ஒரு தசமப் புள்ளியையும் அதன் பின் இரண்டு இலக்கங்களையும் சேர்க்கலாம் (Jurafsky and Martin 2000: 29).

/ $ [0 – 9] + \ . [0 – 9] [0 – 9] /

இந்த அமைப்பொழுங்கு $199.99 என்பதை அனுமதிக்கும் ஆனால் $199 என்பதை அனுமதிக்காது. நாம் செண்ட்களைக் கட்டாயமற்றதாகச் செய்யவேண்டும் மற்றும் நாம் சொல்லின் எல்லையில் இருப்பதாக நிச்சயிக்கவேண்டும் (Jurafsky and Martin 2000: 29).

/ \b $ [0 – 9] + \ . ([0 – 9] [0 – 9]) ? \ b /

எவ்வாறு நாம் செயலியின் வேகத்தைக் (processor speed) குறிப்பிடலாம், megahertz = MHz அல்லது gigahertz = GHz என்ற கேள்வி எழலாம். இதற்கான அமைப்பொழுங்கு கீழே தரப்பட்டுள்ளது (Jurafsky and Martin 2000: 29):

/ \ b [0 – 9] +␣ * (MHz | [Mm] egahertz | GHz | [Gg] igahertz) \ b /

நாம் /␣* / என்பதை "பூஜியம் அல்லது கூடுதல் இடைவெளிகள்" என்பதைக் குறிப்பிடப் பயன்படுத்தலாம்; ஏனென்றால் அதிகப்படியான இடைவெளிகள் சுற்றி இருக்காலாம். தட்டு இடைவெளியை(disk space) (Gb = gigabytesஇல்) அல்லது நினைவக அளவைக் (memory size) (Mb = megabytesஇல் அல்லது Gb = gibabytesஇல்) கையாளுகையில் நாம் கட்டாயமல்லா (optional) கிகாபைட் பின்னங்களை அனுமதிக்க வேண்டிவரும். இறுதி $ கட்டாயமில்லாமையைக் குறிப்பிட ? ஐப் பயன்படுத்துவதைக் குறித்துக்கொள்ளவும் (Jurafsky and Martin 2000: 30).

/ \ b [0 – 9] +␣* (Mb | [Mm] egabytes ?) \ b/

/ \ b 0 – 9 (\ . [0 – 9] +) ? ␣* (Gb | [Gg] igabytes?) \ b/

இறுதியாக நாம் இயக்க ஒழுங்குமுறைகள் (operating systems) மற்றும் வெண்டர்கள் (vendors) இவற்றைக் குறிப்பிடும் எளிய அமைபொழுங்கை வேண்டுவோம்:

/ \ b (Win95 | Win98 | WinNT | Windows␣* (NT | 95 | 98 | 2000) ?) \ b /

2.3.1.5. உயர்நிலை இயக்கிகள் (Advanced Operators)

சில பயனுள்ள மேம்பட்ட வழக்கமான வெளிப்பாடு இயக்கிகள்/ஆபரேட்டர்கள் உள்ளன. கிழே தரப்பட்டுள்ள அட்டவணை பொதுவான வரம்புகளுக்கு சில பயனுள்ள மாற்றுப்பெயர்களைக் காட்டுகின்றது. இவை முக்கியமாகத் தட்டச்சு சிக்கானத்திற்குப் பயனுள்ளவை. க்ளீன் * மற்றும் க்ளீன் + தவிர, நாம் வெளிப்படையான எண்களைச் சுருள் அடைப்புக்குள் (curly brackets) தந்து எண்ணிகளாகப் (counters) பயன்படுத்தலாம். வழக்கமான

வெளிப்பாடு /{3}/ என்பது "முந்தைய எழுத்துக்களின் சரியான 3 நிகழ்வுகள் அல்லது வெளிப்பாடு ". எனவே /a \.{24 }z / என்பது a-ஐத் தொடர்ந்து z-ஆல் தொடரப்பட்ட 24 புள்ளிகளுடன் பொருந்துகிறது (ஆனால் a-ஐத் தொடர்ந்து z-ஆல் தொடரப்பட்ட 23 அல்லது 25 புள்ளிகளுடன் பொருந்தாது) (Jurafsky and Martin 2000: 30)..

குறிப்பிட்ட எண்ணிக்கையிலான எண்களையும் குறிப்பிடலாம்; எனவே / {n, m} / என்பது முன்வரும் எழுத்தோ வெளிப்பாடோ n முதல் m தடவைகள் வருவதைக் குறிப்பிடும். {n} என்பது முன்வரும் எழுத்தோ வெளிப்பாடோ n தடவைகள் வருதைக் குறிப்பிடும்.

இறுதியாக, குறிப்பிட்ட சிறப்பு எழுத்துக்கள் பின்சாய்வு (\) அடிப்படையிலான சிறப்பு குறிப்பால் குறிப்பிடப்படுகின்றன. இவற்றில் மிகவும் பொதுவானது புதியவரி எழுத்து (new line character) \n மற்றும் தாவலின் எழுத்து \t. (., *, [, \ போன்ற) சிறப்பான எழுத்துக்களைக் குறிப்பிட அவற்றின் முன்பு ஒரு பின்சாய்வுக்கோடு பயன்படுத்தப்படுகின்றது (அதாவது / \. /, / \ * /, / \ [/, மற்றும் / \\ /) (Jurafsky and Martin 2000: 30).

2.3.1.6. பொதுவான எழுத்துகளின் குழுமங்களின் மறுகுறியீடுகள்

பொதுவான எழுத்துகளின் குழுமங்களின் மறுகுறியீடுகள் பின்வரும் அட்டவணையில் தரப்பட்டுள்ளன (Jurafsky and Martin 2000: 30):

சீரான வெளிப்பாடுகள்	**விரிவு**	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
\d	[0-9]	எந்த ஸ்தான எண்	Party_ of_ 5
\D	[^0-9]	எந்த ஸ்தான எண் அல்ல	Blue_ moon
\w	[a-zA-Z0-9_]	எந்த எழுத்து-எண் அல்லது இடைவெளி	Daiyu
\W	[^\W]	எந்த எழுத்து-எண்ணும் அல்ல	!!!!
\s	[_\r\t\n\f]	வெள்ளை இடைவெளி (இடைவெளி, தாவல்)	
\S	[^\S]	வெள்ளை இடைவெளி அல்ல	

2.3.1.7. எண்ணுவதற்கான சீரான வெளிப்பாட்டு இயக்கிகள்

பின்வரும் அட்டவணையில் எண்ணுவதற்கான இயக்கிகளும் எடுத்துக்காட்டுகளும் தரப்பட்டுள்ளன (Jurafsky and Martin 2000: 31):

சீரான வெளிப்பாடு	பொருத்தம்
*	முன்வரும் எழுத்து பூஜியமோ அதற்கு மோலோ தடவைகள் வருதல்
+	முன்வரும் எழுத்து ஒன்றோ அதற்கு மேலோ தடவைகள் வருதல்
?	முன்வரும் எழுத்த சரியாக பூஜியமோ ஒரு தடவையோ வருதல்
{n}	முன்வரும் எழுத்து n தடவைகள் வருதல்
{n,m}	முன்வரும் எழுத்து n முதல் m தடவைகள் வருதல்
{n,}	முன்வரும் எழுத்து குறைந்தது n தடவைகள் வருதல்

2.3.1.8. இடச்சாய்கோடு தேவையான சில எழுத்துகள்

இடச்சாய்கோடு தேவையான சில எழுத்துகள் கீழ்வரும் அட்டவணையில் தரப்பட்டுள்ளன (Jurafsky and Martin 2000: 31):

சீரான வெளிப்பாடு	பொருத்தம்	பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு
*	ஆஸ்டரிக்ஸ் நட்சத்திரக் குறியீடு '*'	"K*A*P*L*A*N"
\.	முழுப்புள்ளி "."	"Dr. Livingston"
\?	கேள்விக்குறி "?"	"Should you light my coedle?"
\n	புது வரி	
\t	ஒரு தட்டு	

2.3.1.9. சீரான வெளிப்பாட்டு இடப்பெயர்ப்பு, நினைவகம், மற்றும் எலிசா

இடப்பெயர்ப்புகள் (substitution) சீரான வெளிப்பாடுகளின் முக்கியப் பயன்பாடாகும். எடுத்துக்காட்டாகப் பெர்ல் இடப்பெயர்ப்பு இயக்கி (Perl substitution operator) s / regexp1 / regexp2 / என்பது ஒரு சீரான வெளிப்பாட்டால் பண்பாக்கம் செய்யப்பட்ட எழுத்துக்களின் ஒரு கோர்வையை வேறு ஒரு சீரான வெளிப்பாட்டால் பண்பாக்கம் செய்யப்பட்ட கோர்வையால் இடம்பெயர்ப்பதாகும் (Jurafsky and Martin 2000: 31):

s / colour / color /

ஒரு கோர்வையின் ஒரு குறிப்பிட்ட துணைப் பகுதியைக் குறிப்பிட இயலுவது பயனுள்ளதாகும். எடுத்துக்காட்டாக நாம் கோண அடைப்புக் குறிகளைப் (angle brackets) பயன்படுத்தி முழு எண்களை உட்படுத்தலாம்: எ.கா. the 35 boxes - the <35> boxes. முழு எண்ணைப் பின் குறிப்புரை செய்ய ஒரு வழியை விரும்பலாம்; இதற்குப் பிறை அடைப்புக் குறிகளைச் (parenthesis) சேர்க்கலாம். நாம் முதல் அமைப்பொழுங்கைச் சுற்றி பிறை அடைப்புக் குறிகளை '( )' இடுகின்றோம்; பின் குறிப்புரை செய்ய எண் இயக்கி \1 ஐ இரண்டாம் அமைப்பொழுங்கில் பயன்படுத்துகின்றோம்.

s/ ([0 – 9] +) / < \ 1>/

பிறை அடைப்புக் குறிகளையும் எண் இயக்கிகளையும் ஒரு உரையில் இரு தடவை வரும் குறிப்பிட்ட கோர்வையை அல்லது கூற்றைக் குறிப்பிட பயன்படுத்த இயலும். எடுத்துக்காட்டாக, நாம் தேடும் இரண்டு X உடைய என்பதை ஒரே கோர்வைக்குக் கட்டுப்படுத்த விரும்பும் "the Xer they were, the Xer they will be" என்ற அமைப்பொழுங்கை எடுத்துக்கொள்வோம். நாம் இதைப் பின்வருமாறு முதல் Xஐச் சுற்றி பிறை அடைப்புக்குறி இயக்கியை இட்டு மற்றும் இரண்டாவது Xஐ எண் இயக்கி \1 ஆல் இடம்பெயர்த்து செய்கின்றோம்:

/ the (.*) er they were, the \ 1 er they will be /

இங்கு \ 1 என்பது பிறை அடைப்புக்குறிக்குள் இருக்கும் முதல் உருப்படியுடன் பொருந்தும் எந்தக் கோர்வையாலும் இடம்பெயர்க்கப்படும். எனவே இது *The bigger they were, the bigger they will be* என்பதுடன் பொருந்தும்; *The bigger they were, the faster they will be* என்பதுடன் பொருந்தாது.

எண் இயக்கியைப் பிற எண்களுடன் பயன்படுத்தலாம்: நான் இரண்டு வேறுபட்ட கண அடைபுக்குறிகளைப் பொருத்தினால், \ 2 என்பது என்பது இரண்டாவது கணத்துடன் பொருந்தும் எதுவும் என்று பொருள்படும். எடுத்துக்காட்டாக,

/ the (.*) er they (.*) , the \1 er they \ 2 /

என்பது *The bigger they were, the bigger they were* என்பதுடன் பொருந்தும்; *The bigger they were, the bigger they will be* என்பதுடன் பொருந்தாது. இந்த எண்ணிடப்பட்ட நினைவகங்கள் பதிவுகள் (registers) (எ.கா. பதிவு 1, பதிவு 2, பதிவு 3 போன்றன) என்று அழைக்கப்படுகின்றன. இந்த நினைவகப் பண்புக்கூறு ஒவ்வொரு சீரான வெளிப்பாட்டு மொழியின் பகுதி அல்ல; இது பெரும்பான்மையும் சீரான வெளிப்பாடுகளின் "நீட்சிசெய்யப்பட்ட" பண்புக்கூறாகக் கருதப்படுகின்றது.

நினைவகத்தைப் பயன்படுத்திய இடப்பெயர்பெயர்ப்புகள் எலிசா (**ELIZA)** (Weizembaum, 1966) போன்ற எளிய இயற்கை மொழி புரிதல் வழிமுறை வரைவைப் (natural-language understanding program) பிரயோகிப்பதற்குப் மிகவும் பயனுள்ளதாகும் (Jurafsky and Martin 2000: 33)

. ரோகெரியன் உளவியலார் **ELIZA**ஆல் தூண்டப்பட்ட பின்வரும் உரையாடலை நினைவூட்டிக் கொள்ளவும்.

User1: Men are all alike.

ELIZA1: IN WHAT WAY.

User2: They're always bugging us about something or other.

ELIZA2: CAN YOU THINK OF A SPECIFIC EXAMPLE.

User3: Well, my boyfriend made me come here.

EIZA3: YOUR BOYFRIEND MADE YOU COME HERE

User4: He says I'am depressed much of the time.

ELIZA4: I AM SORRY TO HEAR YOU ARE DEPRESSED.

உள்ளீடு செய்யப்பட்ட சீரான வெளிப்பாட்டு இடப்பெயர்ப்புகளின் அடுக்குகளைக் கொண்டு உள்ளீடு செய்யப்பட்ட வரிகளின் எதாவது பகுதியை பொருத்தம் செய்து மாற்றி (இடம்பெயர்த்து) எலிசா செயல்படும். எடுத்துக்காட்டாக முதலாவது இடப்பெயப்புகள் *my* என்பதன் எல்லா நேர்வுகளையும் *YOUR* என்பதாலும் *I'am* என்பதை *YOU ARE* என்பதாலும் மாற்றம் செய்துள்ளது. அடுத்த கணங்களின் இடப்பெயர்ப்புகள் உள்ளீட்டில் தொடர்புடைய அமைப்பொழுங்குகளைத் தேடி பொருத்தமான வெளியீட்டை உருவாக்கும். கீழே சில எடுத்துக்காட்டுகள் தரப்பட்டுள்ளன:

s /.* YOU ARE (depressed | sad) .* / I AM SORRYTO HEAR YOU ARE \1 /

s /.* YOU ARE (depressed | sad) .* / WHY DO THINK YOU ARE \1 /

s/ .* all .* / IN WHAT WAY /

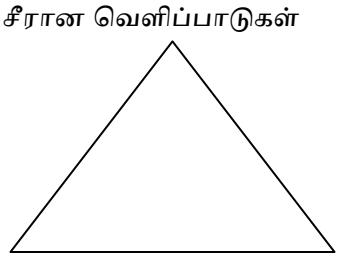
s / .* always .* / CAN YOU THINK OF A SPECIFIC EXAMPLE /

சில பன்மடங்கான இடப்பெயர்ப்புகளைத் தரப்பட்ட உள்ளீட்டுக்குப் பிரயோகிக்க இயலும் என்பதால் இடப்பெயர்ப்புகளுக்குத் தர எண்கள் தரப்பட்டுள்ளன; அந்த நிரலில் அவை பிரயோகிக்கப்படும்.

2.4.2. முற்றுநிலைத் தானியங்கி (Finite State Automata FSA)

சீரான வெளிப்பாடு பனுவல் தேடலுக்கு வசதியான கருவிமொழி என்பதிலும் மேற்பட்டதாகும். முதலாவதாகச் சீரான வெளிப்பாடு ஒரு முற்றுநிலைத் தானியங்கியை விளக்குவதற்கான ஒரு வழியாகும் (Jurafsky and Martin 2000: 33-53). இறுதிநிலைத் தானியங்கிகள் கணிக்கும் வேலையின் முக்கிய காரியங்களின் கோட்பாடு அடித்தளமாகும். (நினைவுப் பண்புக்கூறைப் பயன்படுத்தும் சீரான வெளிப்பாடு தவிர) எந்த சீரான வெளிப்பாட்டையும் இறுதிநிலைத் தானியங்கிகளாக நிறைவேற்றலாம். அதே போல் எந்த முற்றுநிலைத் தானியங்கியையும் சீரான வெளிப்பாட்டால் விளக்கலாம். இரண்டாவதாக சீரான வெளிப்பாடு என்பது சீரான மொழி (Regular Language) என்று அழைக்கப்படுகின்ற ஒரு வகைக் குறிப்பிட்ட முறையான மொழியைப் (Formal Language) பண்பாக்கம் செய்யும் ஒரு வழியாகும். சீரான வெளிப்பாடுகளையும் இறுதிநிலை தானியங்கிகளையும் முறையான மொழியை விளக்கப் பயன்படுத்தலாம். இந்த மூன்று கோட்பாட்டுக் கட்டமைப்புகளுக்கு (three theoretical constructions) இடையே உள்ள தொடர்பைப் பின்வரும் படத்தால் காட்டலாம் (Jurafsky and Martin 2000: 33).

படம் 3:



முற்றுநிலைத் தானியங்கி சீரான வெளிப்பாடுகளை நிறைவேற்றப் பயன்படுத்தப்படும் ஒரு கணித உபாயம் (device) மட்டுமன்றி கணித மொழியியலின் முக்கியக் கருவிகளில் ஒன்றாகும். தானியங்கிகளின் வேறுபாடுகளான முற்றுநிலை மாறிகள் (Finite State transducers (FST)), ஹிடன் மார்க்கவ் மாதிரிகள் (Hidden Markov Models (HMM)) மற்றும் N கிராம் இலக்கணங்கள் (N.Gram grammers) என்பன பேச்சைத் தெரிந்து கொள்ளுதல் (speech recognition), பேச்சு உருவாக்கம் (speech synthesis), எழுத்துத் திருத்தம் (spell checking), செய்திப் பிரித்தெடுப்புப் பயன்பாடுகள் (Information Extraction Applications) என்பனவற்றின் முக்கியமான பகுதிகளாகும்.

2.4.2.1. முற்று நிலைத் தானியங்கிகளைக் கொண்டு ஆட்டுமொழியைப் (Sheep talk) புரிந்து கொள்ளுதல்

முற்றுநிலைத் தானியங்கிகளின் செயல்பாட்டை விளக்க ஆட்டுமொழியின் உருப்படுத்தத்தைப் பயன்படுத்தலாம் (Jurafsky and Martin 2000: 34-37). ஆட்டு மொழியில் பின்வரும் குழுமத்தில் உள்ள ஏதாவது ஒரு கோர்வை இருப்பதாகக் கருதலாம்.

ba!

baa!

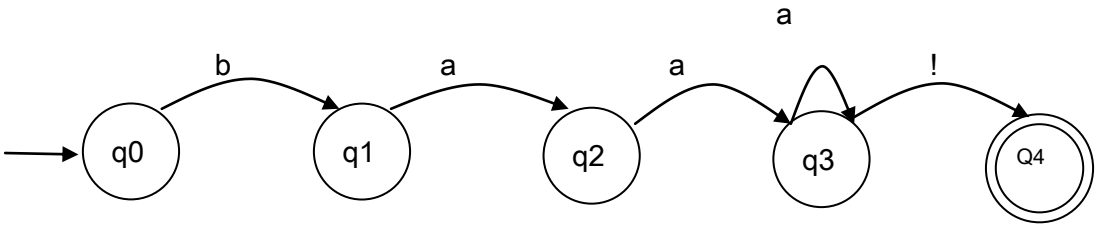
baaa!

baaaa!

baaaaa!

இம்மாதிரியான ஆட்டுமொழியின் சீரான வெளிப்பாடு /baa+!/ என்று அமையும். பின்வரும் படம் இந்த சீரான வெளிப்பாட்டின் முற்று நிலை தானியங்கி மாதிரியாக அமையும்.

படம் 4:



முற்று நிலைத் தானியங்கி ஒரு குழுமம் கோர்வைகளைப் புரிந்துக் கொள்கிறது. இங்குச் சீரான வெளிப்பாடு செய்வது போன்று ஆட்டுமொழியைப் பண்பாக்கம் செய்யும் கோர்வைகளைப் புரிந்து கொள்கிறது. கணுக்களின் முற்றுக் குழுமங்கள் (finite vertices/nodes) மற்றும் இரு கணுக்களை இணைக்கும் விற்கள் மூலமும் நாம் தானியங்கியைத் திசைநோக்கு வரைபடமாக (directed graph) உருப்படுத்தம் செய்கிறோம். நாம் கணுக்களை வட்டங்களாகவும் விற்களை அம்புக் குறிகளாலும் உருப்படுத்தம் செயகிறோம். தானியங்கிக்கு ஐந்து நிலைகள் உண்டு. அவை வரைப்படத்தில் கணுக்களாக உருப்படுத்தம் செய்யப்பட்டுடள்ளன. பூஜியநிலை தொடக்க நிலையாகும். இது உள்வரு அம்புக்குறியாக உருப்படுத்தம் செய்யயப்பட்டுள்ளது. நாலாவது நிலை இறுதிநிலை அல்லது ஏற்கும் நிலை (accepting strate) ஆகும். இது இரட்டை வட்டத்தால் (double cricle) உருப்படுத்தம் செய்யப்பட்டுள்ளது. இதில் நான்கு மாற்றங்கள் உண்டு. அதை நாம் வரைப்படத்தில் விற்களால் உருப்படுத்தம் செய்துள்ளோம். இறுதித் தானியங்கியைக் கோர்வைகளைப் புரிந்து கொள்ள பின்வருமாறு பயன்படுத்தலாம். கட்டங்களாகப் பிரிக்கப்பட்டுள்ள ஒரு நீளமான நாடாவில் கட்டத்திற்கு ஒரு குறியீடாக எழுதப்படுவதாகக் கொள்ளலாம். தானியங்கி q_0 என்ற தொடக்க நிலையில் தொடங்குகின்றது. பின்வரும் செயல்பாட்டைத் திரும்பத்திரும்பச் செய்கின்றது. உள்ளீட்டின் அடுத்த எழுத்தைப் பரிசோதனை செய்கின்றது. அது வில் குறியின் மேலுள்ள குறியீட்டுடன் பொருத்த முற்றால் தற்போதைய நிலையிலிருந்து வில்லைத் தாண்டி அடுத்த நிலைக்குச் செல்கிறது. மட்டுமல்லாமல் உள்ளீடு

இல்லாத ஏற்பு நிலையில் (acceping state q_4) இருந்தால் பொறி வெற்றிகரமாக ஆட்டுமொழியின் ஒரு நிகழ்வைப் புரிந்துவிட்டது என்று அர்த்தம். உள்ளீடு இல்லாமலோ வில்லின் மேலுள்ள குறியீட்டுடன் பொருத்தமுறாமலோ பொறி இறுதிநிலைக்குச் செல்லவில்லை என்றாலோ இறுதிக்கு முன்பான நிலையில் தடங்கல் ஏற்பட்டது என்றாலோ நாம் பொறி உள்ளீட்டை தள்ளிவிட்டது என்றோ ஏற்றுக் கொள்ளவில்லை என்றோ கூறலாம். நாம் தானியங்கியை நிலைமாற்றமைவு அட்டவணையாலும் (state-transition table) உருப்படுத்தம் செய்யலாம். பின்வருவது மேற்கண்ட இறுதிநிலைத் தானியங்கியின் நிலைமாற்ற அட்டவணையாகும்.

c	உள்ளீடு
நிலை	ba !
0	1 Ø Ø
1	Ø 2 Ø
2	Ø 3 Ø
3	Ø 3 4
4	Ø Ø Ø

முக்காற் புள்ளியுடன் (:) கூடிய 4-வது நிலை இறுதிநிலையைக் குறிக்கும். Ø தவறான அல்லது விட்டுப்போன மாற்றத்தைக் குறிக்கும். ஒரு இறுதிநிலைத் தானியங்கியைப் பின்வரும் 5 அளவுகளால் (parameters) விளக்கலாம் (Jurafsky and Martin 2000: 36).

Q: N நிலைகளின் முற்றுக் குழுமம், q_0, q_1, … q_N

Σ: குறியீடுகளின் முற்று உள்ளீட்டு அகர எழுத்து

q_0 : தொடக்கநிலை

F : இறுதி நிலைகளின் குழுமம் $F \subseteq Q$ ·

δ(q,i):நிலைகளுக்கு இடையேயுள்ள மாற்றமைவுச் செயல்பாடு அல்லது மாற்றமைவு வார்ப்புரு (Matrix) $q \in E$ மற்றும் உள்ளீட்டுக்குறியீடு $i \in \Sigma$, δ (q,i) · என்று தரப்பட δ (q,i) என்பது $q' \in Q$ என்ற புதிய நிலைக்குத் திரும்பும். இவ்வாறு δ என்பது Q X Σ இலிருந்து Q –உக்கு உள்ள தொடர்பு.

மேற்கண்ட படத்தில் உள்ள ஆட்டு மொழித் தானியங்கிக்கு

$Q = \{ q_0, q_1, q_2, q_3, q_4 \}$

$\Sigma = \{a, b, !\}$

$F = \{q_4\}$

δ (q, i) மேற்கண்ட மாற்றமைவு அட்டவணையால் விளக்கப்பட்டுள்ளது. கீழே தரப்பட்டுள்ள நிலைமாற்றமைவு அட்டவணையைப் பயன்படுத்திய ஒரு கோர்வையைப் புரிந்துகொள்வதற்கான வழிமுறை வரைவு (Jurafsky and Martin 2000: 37).

செயல்பாடு உறுதிசெய்யப்பட்ட புரிந்துகொள்கை (நாடா, பொறி) ஏற்றுக்கொள் அல்லது விட்டுவிடு

வரிசைகிரமம் (index) → நாடாவின் தொடக்கம்

நிகழ்நிலை → பொறியின் ஆரம்பநிலை

கண்ணி (Loop)

உள்ளீட்டின் இறுதி நிலை எட்டிவிட்டால் பின்னர்
நிகழ் நிலை ஏற்பு நிலை என்றால் பின்னர்
ஏற்றுக் கொள்.
விட்டு விடு

இல்லை என்றால் (elsif) மாற்றமைவு அட்டவணை [நிகழ் நிலை, நாடா [வரிசை] காலி என்றால் பின்னர்
விட்டு விடு.
மாறாக

நிகழ் நிலை ←மாற்றமைவு அட்டவணை [நிகழ் நிலை, நாடா[வரிசை கிரமம்]]

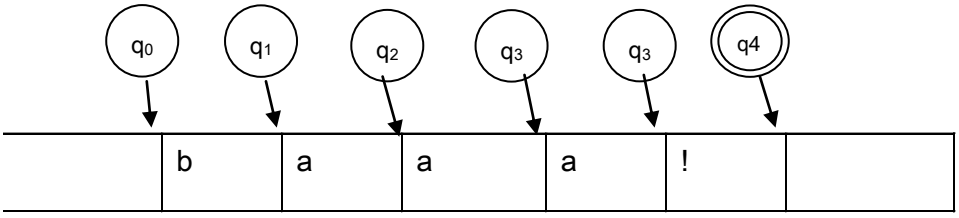
வரிசை கிரமம் ← வரிசை கிரமம் +1

முடிவு

உறுதி செய்யப்பட்ட புரிந்துகொள்கை (D-RECOGNIZE) ஒரு நாடாவையும் ஒரு தானியங்கியையும் உள்ளீடாக ஏற்கிறது. அது நாடாவில் சுட்டுகிற கோர்வை தானியங்கியால் ஏற்றுக் கொள்ளப்பட்டால் அது ஏற்புக்குத் திரும்புகிறது. இல்லாவிடில் தள்ளிவிடுகிறது. உறுதி செய்யப்பட்ட புரிந்துகொள்கையானது நாடாவின் தொடக்கத்தில் உள்ள மாறி வரிசைக்கிரமம் மற்றும் பொறியின் தொடக்கநிலையின் தற்போதைய நிலையைத் தூண்டித் தொடங்குகிறது. உறுதிசெய்யப்பட்ட புரிந்துகொள்கை பின்னர் வழிமுறை வரவின் மீதமுள்ளதைச் செயல்படுத்தும் கண்ணிக்குள் செல்கிறது. அது முதலில் அதன் உள்ளீட்டின் இறுதியை அடைந்து விட்டதா என்று பரிசீலிக்கின்றது. அப்படி என்றால் (நிகழ்நிலை ஒரு ஏற்புநிலை என்றால்) அது உள்ளீட்டை ஏற்றுக் கொள்ளும் அல்லாவிடில் உள்ளீட்டைத் தள்ளிவிடும். நாடாவில் உள்ளீடு மீதமிருந்தால் உறுதிசெய்யப்பட்ட தெரிகை எந்த நிலைக்குச் செல்ல வேண்டும் என்பதைத் தீர்மானிக்க மாற்றமைவு அட்டவணையைப் பார்க்கிறது. மாறியான நிகழ்நிலை அட்டவணையிலுள்ள எந்தச் செங்குத்து வரிசையைப் பார்க்க வேண்டும் என்று காட்டும். நாடாவிலுள்ள நிகழ் குறியீடு (current symbol) அட்டவணையிலுள்ள எந்த கிடக்கை வரிசையை பார்க்க வேண்டும் என்பதைக் காட்டும். விளையும் மாற்றமைவு அட்டவணைக் கட்டம் மாறியான தற்போதைய நிலையைப் (current-state) புதுப்பிக்கப் பயன்படுத்தப்படுகின்றது; வரிசைக் கிரமம் (index) நாடாவில் முன்செல்ல

உந்தப்படுகின்றது. மாற்றமைவு அட்டவணைக் கட்டம் காலியாக இருந்தால் பொறிக்கு எங்கும் போக இயலாதாகையால் அது உள்ளீட்டைத் தள்ளிவிடுகின்றது. பின்வரும் படம் தரப்பட்ட baaa! என்ற உள்ளீட்டுக் கோர்வையான ஆட்டின் மொழி முற்றுநிலை தானியங்கி மேல் மேற்சொன்ன வழிமுறை வரையின் நிறைவேற்றலைக் காட்டுகின்றது (Jurafsky and Martin 2000: 37).

படம் 6:



நாடாவின் தொடக்கத்தைப் பரிசோதிப்பதற்கு முன் பொறி q_0 என்ற நிலையில் இருக்கிறது. உள்ளீட்டு நாடாவில் b-ஐ கண்டுகொண்டதும் அது மாற்றமைவு அட்டவணையின் உள்ளடக்கத்தில் [q_0,b] சுட்டப்பட்டது போல் q_1 நிலைக்கு மாறுகின்றது. அது பின்னர் a-ஐக் கண்டுகொண்டு q_2 நிலைக்கு மாறுகின்றது. மற்றொரு a அதை q_3 நிலைக்குக் கொண்டு செல்கின்றது. முன்றாவது a அதை q_3 நிலையில் விட்டுவிடுகின்றது. அங்கு அது "!" என்பதைக் கண்டுகொண்டு q_4 நிலைக்குச் செல்கின்றது. வேறு உள்ளீடு இல்லாத காரணத்தால் கண்ணியின் தொடக்கத்தில் வரும் "உள்ளீட்டின் இறுதி (end of input)" என்ற நிபந்தனை முதல் தடவையாக பூர்த்தியாக்கப்பட்டு பொறி q_4 நிலையில் நிற்கும். q_4 என்ற நிலை ஏற்பு நிலை எனவே பொறியானது baaa! என்ற கோர்வையை ஆட்டுமொழியில் உள்ள ஒரு வாக்கியமாக ஏற்றுக்கொண்டது.

2.4.2.2. முறையான மொழிகள் (Formal Languages) (Jurafsky and Martin 2000: 38-39)

பின்வரும் வரைபடத்தை ஆட்டின் மொழியை ஆக்கும் தானியங்கியாகப் பயன்படுத்தலாம். இதற்கு நாம் தானியங்கி q_0 நிலையில் தொடங்கி விற்களைத் தாண்டி புதிய நிலைகளை அடைந்து ஒவ்வொரு வில்லிலும் உள்ள குறியீடுகளை அச்சிடுவதாகக் கொள்ள வேண்டும். தானியங்கி இறுதி நிலையை அடைந்ததும் நிற்கும். நிலை 3-இல் தானியங்கி a ! என்பதை அச்சிட்டு நிலை 4-க்கு செல்வதையோ அல்லது a-ஐ அச்சிட்டு விட்டு நிலை 3-க்கு திரும்புவதை தெரிந்தெடுக்க வேண்டும். கோர்வை முறையான வெளிப்பாட்டால் கண்டுபிடிக்கப்பட்ட ஆட்டுமொழிக்கான கோர்வையாக இருப்பது வரை, தற்போது பொறி எவ்வாறு இந்தத் தீர்மானத்தைச் செய்கிறது என்றும்

ஆட்டுமொழியின் எந்த கோர்வையை ஆக்குகின்றது என்பது குறித்தும் கவனத்தில் கொள்ள வேண்டியதில்லை.

ஒரு முறையான மொழியின் எல்லா மற்றும் அதற்குரிய கோர்வைகளை மட்டும் ஆக்கம் செய்யும் மற்றும் தெரிந்துகொள்ளும் மாதிரிதான் முறையான மொழியின் விளக்கமாகச் செயல்படும். ஒரு முறையான மொழி என்பது கோர்வைகளின் ஒரு கணம் ஆகும்; ஒவ்வொரு கோர்வைகளும் அகரவரிசை எழுத்து என்று அழைக்கப்படுகின்ற ஒரு முற்று குறியீட்டு கணத்திலிருந்து வரும் குறியீடுகளால் ஆனது. ஆட்டின் மொழியின் அகர வரிசை எழுத்து Σ ={ a,b,!} என்ற கணம் ஆகும். m என்ற மாதிரி தரப்படுகையில் நாம் L(m) என்பதை "m என்பதால் பண்பாக்கம் செய்யப்பட்ட முறையான மொழி" என்று பொருள்பட பயன்படுத்தலாம். எனவே நாம் முன்னர்பார்த்த ஆட்டின் மொழி தானியங்கியால் விளக்கப்பட்ட முறையான மொழியை பின் வருமாறு முற்றுக் கணமாகக் கொள்ளலாம்.

L(m) = {baa!, baaa!, baaaa!, baaaaa!,...}

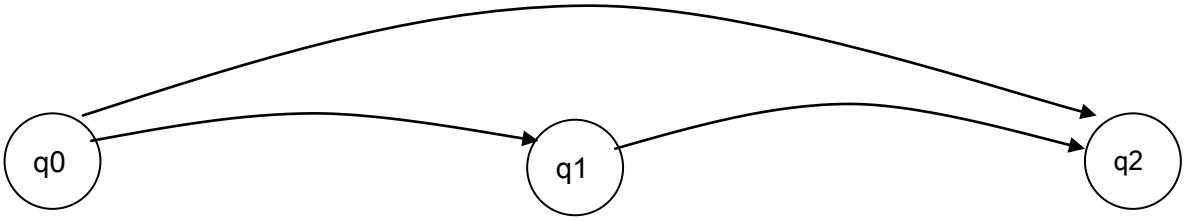
ஒரு மூடப்பட்ட வடிவில் முற்றிலாக கணத்தை வெளியிடலாம் என்பதுதான் ஒரு மொழியை விளக்க தானியங்கியின் பயன்பாடாகும். முறையான மொழிகள் மக்கள் பேசும் இயற்கை மொழிகள் (natural language) அல்ல. முறையான மொழி எந்த வகையிலும் உண்மையான மொழியின் ஒப்புமை கொள்ளாமலிருக்கலாம் (எடுத்துக்காட்டாக ஒரு முறையான மொழியை ஒரு சோடா இயந்திரத்தின் வேறுபட்ட நிலைகளை மாதிரிப்படுத்தப் பயன்படுத்த இயலும்). இருப்பினும் நாம் முறையான மொழியை ஒலியனியல், உருபனியல், தொடரியல் போன்ற இயற்கை மொழியின் பகுதிகளை மாதிரிப்படுத்தப் பயன்படுத்துகின்றோம். ஆக்கமுறை இலக்கணம் (generative grammar) என்ற கலைச்சொல் ஒரு முறையான மொழியின் இலக்கணத்தைக் குறிப்பிடுமாறு பயன்படுத்தப்படுகின்றது: எல்லா இயலக்கூடிய கோர்வைகளை ஆக்கி ஒரு மொழியை விளக்க ஒரு தானியங்கியாகப் பயன்படுவது இந்தக் கலைச் சொல்லின் மூலமாகும்.

2.4.2.3 மற்றொரு எடுத்துக்காட்டு

முன்னர் கூறிய எடுத்துக்காட்டுகளில் முறையான அகரவரிசை/நெடுங்கணக்கு எழுத்துக்களால் ஆனது. ஆனால் சொற்களால் ஆன உயர்நிலையிலும் இதை அமல்படுத்தலாம். இவ்வழியில் நாம் சொற்சேர்க்கையைப் பற்றிய உண்மைகளை மாதிரிப்படுத்திய முற்றுநிலை தானியங்கிகளை எழுதலாம். அம்மாதிரியான முறையான மொழி பத்துபைசா, மூன்று ரூபாய், ஒருரூபாய் முப்பது பைசா போன்ற தொடர்களால் ஆன தமிழ்த் துணைக் கணத்தை (subset) மாதிரிப்படுத்தும்.

படம் 7:

ஒன்று	ஆறு	பதினொன்று	பதினாறு	இருபது	எழுபது
இரண்டு	ஏழு	பதுமூன்று	பதினேழு	முப்பது	எண்பது
மூன்று	எட்டு	பதிநான்கு	பதினெட்டு	நாற்பது	தொண்ணூறு
நான்கு	ஒன்பது	பதினைந்து	பத்தொன்பது	ஐம்பது	
ஐந்து	பத்து			அறுபது	

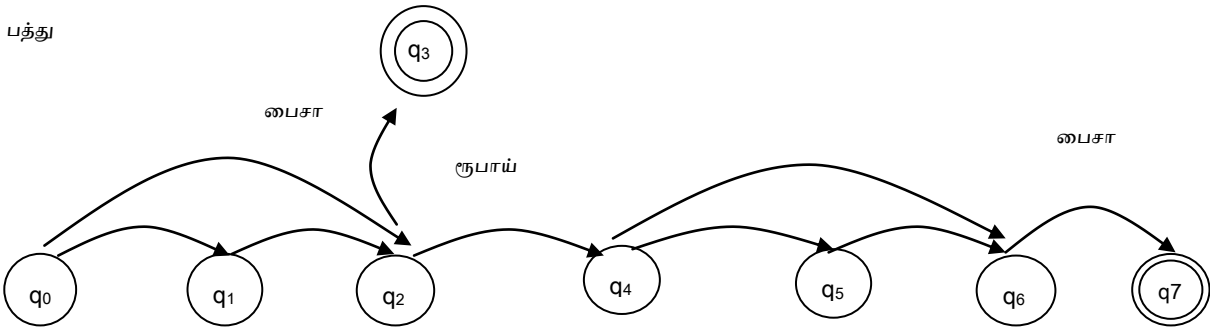


இருபது	அறுபது		ஒன்று	ஆறு	
முப்பது	எழுபது		இரண்டு	ஏழு	
நாற்பது	எண்பது		மூன்று	எட்டு	
ஐம்பது	தொண்ணூறு		நான்கு	ஒன்பது	
			ஐந்து		

படம் 8:

ஒன்று/	பதினொன்று	இருபது
ஒரு	பதுமூன்று	முப்பது
இரண்டு	பதிநான்கு	நாற்பது
மூன்று	பதினைந்து	ஐம்பது
நான்கு	பதினாறு	அறுபது
ஐந்து	பதினேழு	எழுபது
ஆறு	பதினெட்டு	எண்பது
ஏழு	பத்தொன்பது	தொண்ணூறு
எட்டு		
ஒன்பது		
பத்து		

ஒன்று/ஒரு	பதினொன்று	இருபது
இரண்டு	பதுமூன்று	முப்பது
மூன்று	பதிநான்கு	நாற்பது
நான்கு	பதினைந்து	ஐம்பது
ஐந்து	பதினாறு	அறுபது
ஆறு	பதினேழு	எழுபது
ஏழு	பதினெட்டு	எண்பது
எட்டு	பத்தொன்பது	தொண்ணூறு
ஒன்பது		
பத்து		



இருபது	ஒன்று/ஒரு
முப்பது	இரண்டு
நாற்பது	மூன்று
ஐம்பது	நான்கு
அறுபது	ஐந்து
எழுபது	ஆறு
எண்பது	ஏழு
தொண்ணூறு	எட்டு
	ஒன்பது

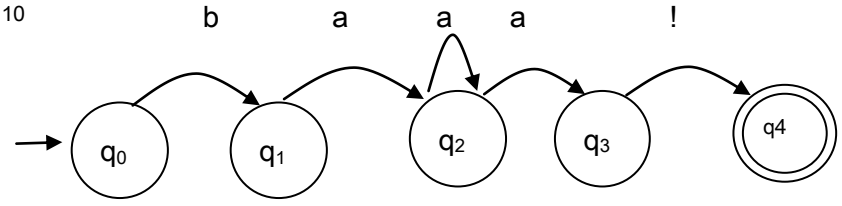
இருபது	ஒன்று/ஒரு
முப்பது	இரண்டு
நாற்பது	மூன்று
ஐம்பது	நான்கு
அறுபது	ஐந்து
எழுபது	ஆறு
எண்பது	ஏழு
தொண்ணூறு	எட்டு
	ஒன்பது

நாம் ரூபாயின் வேறுபட்ட (நூறு, ஆயிரம், போன்ற) எண்ணிக்கைகளுக்கு இலக்கணத்தில் மேலும் சில சொற்கூறுகளைச் சேர்க்க வேண்டி வரும்.

2.4.2.4 . உறுதி செய்யப்படாத முற்றுநிலைத் தானியங்கிகள் (Non-determinisitic Finite State Automatas (NFSAs))

நாம் முன்பு பார்த்த உறுதி செய்யப்பட்ட முற்றுநிலைத் தானியங்கிக்கும் உறுதி செய்யப்படாத தானியங்கிக்கும் உள்ள வேறுபாடு என்னவென்றால் முதல் வகைத் தானியங்கியில் கண்ணி (node) நிலை 2-இலும் தற்போதையத் தானியங்கியில் கண்ணிநிலை 3- இலும் உள்ளது (Jurafsky and Martin 2000: 41).

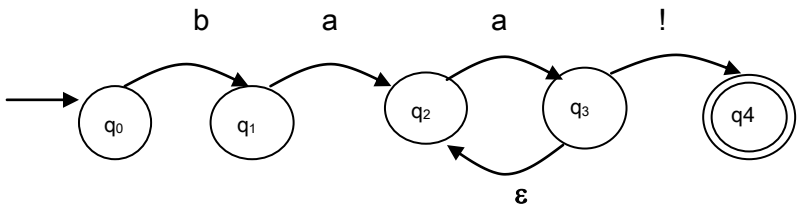
படம் 10



நாம் 2-ஆம் நிலையை அடைந்து ஒரு 'a' ஐப் பார்த்ததும் நிலை 2-இல் இருக்க வேண்டுமா அல்லது 3-ஆம் நிலைக்கு செல்ல வேண்டுமா என்று தெரியவில்லை. இம்மாதிரியான நிச்சயிக்கப்பட வேண்டிய முறைகள் உள்ள தானியங்கிகள் உறுதி செய்யப்படாத முற்றுநிலைத் தானியங்கி (Non- determinisitic Finite State Automate (NFSA)) என்று அழைக்கப்படும்.

குறியீடு இல்லாத விற்களால் ஆன ε மாற்றமைவு (ε-transition) என்று அழைக்கப்படுகின்ற வேறு ஒரு பொது வகை உறுதிசெய்யப்படாத தன்மையும் இருக்கிறது. பின்வரும் படத்திலுள்ள தானியங்கி முன்பு பார்த்த அதே மொழியை இறுதியான ஒன்றாகவோ நமது முதலாவதான ஒன்றாகவோ விளக்குகிறது; ஆனால் இது அதை ε-மாற்றமைவு மூலம் செய்கிறது (Jurafsky and Martin 2000: 41).

படம் 10



இந்தப் புதியவில்லை பின்வருமாறு பொருள் கொள்ளலாம்: நாம் நிலை 3-இல் இருந்தால் நாம் உள்ளீட்டைப் பார்க்காமால் அல்லது நமது உள்ளீட்டுச் சுட்டை முன்னேறாமல் நிலை 2-க்கு செல்ல அனுமதி உண்டு.

2.4.2.5. உறுதிசெய்யப்படாத முற்றுநிலைத் தானியங்கியைக் கோர்வையை ஏற்கப் பயன்படுத்தல் (using an NFSA to accept strings)

நாம் உறுதிசெய்யப்படாத முற்றுநிலைத் தானியங்கியைப் பயன்படுத்தி ஒரு கோர்வை ஆட்டின் மொழியா அல்லவா என்பதை கண்டுகொள்ள, நாம் தவறான வில்லைப் பயன்படுத்தி பின்னர் ஏற்கும் போது விட்டுவிட வேண்டும். ஒரு இடத்தில் ஒன்றிற்கும் மேல் தெரிவு செய்யச் சந்தர்ப்பம் இருப்பதால், நாம் தவறானதை தேர்ந்தெடுக்க இயலும். இந்த உறுதி செய்யப்படாத முற்றுநிலைத் தானியங்கி மாதிரிகளின் சிக்கல்களை நாம் கணினி மாதிரி உருவாக்குகையில் எல்லாம் திரும்பத் திரும்ப நேரிடக்கூடும். இந்தச் சிக்கலுக்கு மூன்று நிலைபெறு தீர்வுகள் இருக்கின்றன (Jurafsky and Martin 2000: 42):

1. காப்பு (backup): எப்பொழுதெல்லாம் தெரிவு இடத்திற்கு (choice point) வருகின்றோமோ நம் உள்ளீட்டில் எங்கிருக்கின்றோம் எந்த நிலையில் தானியங்கி நிற்கிறது என்பதைக் குறிக்க ஒரு அடையாளம் இட இயலும். பின்னர் நாம் தெரிவு செய்தது தவறான வழி என்று புரிந்து கொண்டால் நாம் திரும்பி வந்து வேறு ஒரு பாதையை எடுக்க இயலும்.

2. முன்னோக்கிப் பார்த்தல் (look ahead): நாம் உள்ளீட்டில் முன்னோக்கிப் பார்த்து எந்தப் பாதையை எடுப்பது என்பதைத் தீர்மானிக்க உதவியை நாடலாம்.

3. இணைப்போக்கு **(parallelism):** எப்பொழுதெல்லாம் நாம் தெரிவு இடத்திற்கு (choice point) வருகின்றோமோ நாம் இணையாக இருக்கிற ஒவ்வொரு மாற்றுப் பாதையும் பார்க்கலாம்.

கீழ்வருவது ஆட்டின் மொழிக்கான உறுதிசெய்யப்படாத மாற்றமைவு அட்டவணை (Jurafsky and Martin 2000: 43).

ஆயப்படாத மாற்று விருப்பத்தேர்வுகளுக்கு நாம் எப்போதும் திரும்ப இயலும் என்பதை அறிந்துகொண்டு இறுதிமுனையைக்குக் கொண்டுசெல்லும் விருப்பத்தேர்வுகளை மகிழ்ச்சியுடன்

நாம் செய்யவேண்டும் என்பதை காப்பு அணுகுமுறை பரிந்துரைக்கின்றது. இதற்கு இரண்டு முக்கியக்குறிப்புகள் இருக்கின்றது: ஒவ்வொரு விருப்பத்தேர்வு புள்ளியிடங்களுக்கும் உரிய எல்லா மாற்றுகளையும் நாம் நினைவுகூரவேண்டும் மற்றும் நாம் திரும்ப வர இயலுவதற்காக ஒவ்வொரு மாற்று பற்றியும் போதுமான தகவல்களை நாம் சேகரித்து வைக்கவேண்டும். காப்பு வழிமுறையானது அதன் செயலாக்க முன்னேற்றத்தில் (உள்ளீடு தீர்ந்துபோவதன் காரணத்தாலோ விதிப்படியான நிலைமாற்றங்கள் இல்லாததினாலோ) முன்னேற்றத்தை மேற்கொள்ள இயலாத ஒரு புள்ளியிடத்தை அடையும் போது அது முந்தைய விருப்பத்தேர்வுப் புள்ளியிடத்திற்குத் திரும்புகின்றது. இந்தக் கருத்துச்சாயலை உறுதிசெய்யப்படாத தெரிந்துகொள்வானுக்குப் (non-diterministic recognizer) பிரயோகிக்கையில் நாம் ஒவ்வொரு விருப்பதேர்வுப் புள்ளியிடத்திற்கும் இரண்டு விஷயங்களை நாம் நினைவில்கொள்ளவேண்டும்: நாம் செல்ல இயலும் இயந்திரத்தின் நிலை அல்லது கணு மற்றும் நாடாவின் மீது அதன் தொடர்புடைய இருப்பிடம். கணு மற்றும் இருப்பிடம் இவற்றின் இணைப்பை நாம் தெரிந்துகொள்ளும் வழிமுறையின் தேடல்-நிலை என்று அழைக்கின்றோம். குழப்பத்தைத் தவிர்ப்பதற்காக, நாம் தானியங்கியின் நிலையைக் கணு (node) அல்லது இயந்திர-நிலை (machine state) என்று குறிப்பிடுகின்றோம். இந்த அணுகுமுறை அடிப்படையிலான தெரிந்துகொள்கை வழிமுறைவரைவைப் பின்வருவது அறிமுகப்படுத்துகின்றது (Jurafsky and Martin 2000: 44)

வழிமுறைவரைவு 1

செயல்பாடு உறுதிசெய்யப்படாத தெரிந்துகொள்கை (நாடா, இயந்திரம்) திருப்புகிறது ஏற்றுகொள் அல்லது நிராகரி

நிகழ்ச்சிநிரல் ←{(இயந்திரத்தின் தொடக்க நிலை, நாடாவின் தொடக்கம்)}

தற்போதைய-தேடல்-நிலை←அடுத்தது (நிகழ்ச்சிநிரல்)

கண்ணி

ஏற்றுக்கொள்-நிலை **என்றால்**? (தற்போதைய-தேடல்-நிலை) திருப்புகிறது உண்மை

பின்னர்

திருப்புகிறது ஏற்றுக்கொள்

இல்லாவிடில்

நிகழ்ச்சிநிரல்← நிகழ்ச்சிநிரல் ∪ உருவாக்கு-புதிய-நிலைகள் (தற்போதைய-தேடல்-நிலை)

நிகழ்ச்சிநிரல் காலி **என்றால்** பின்னர்

திருப்புகிறது நிராகரி

இல்லாவிடில்

தற்போதைய-தேடல்- நிலை← அடுத்தது (நிகழ்ச்சிநிரல்)

இறுதி

செயல்பாடு உருவாக்கு-புதிய-நிலைகள்(தற்போதைய-நிலை) திருப்புகிறது தேடல் நிலைகளின் கணம்

தற்போதைய-கணு ← தற்போதைய தேடல்-நிலை இருக்கிற கண்ணி

அகரவரிசை ← தற்போதைய தேடல்-நிலை எதிர்பார்க்கிற நாடாவின் மீது புள்ளியிடம்

திருப்புகிறது பின்வருமாறு மாற்று அட்டவணைகளிலிருந்து ஒரு பட்டியல் தேடல் நிலைகள்:

(மாற்று-அட்டவணை [தற்போதைய-கணு,ε] அகரவரிசை)

∪

மாற்று-அட்டவனை[தற்போதைய-கணு, நாடா[அகரவரிசை]] , அகரவரிசை +1)

செயல்பாடு ஏற்புநிலை ? (தேடல்-நிலை) திருப்புகிறது உண்மை அல்லது பொய்

தற்போதைய-கணு ← தேடல் நிலை இருக்கிற கணு

அகரவரிசை ← தேடல்-நிலை எதிர்பார்க்கிற நாடாவின் மீது புள்ளியிடம்

அகரவரிசை நாடாவின் இறுதியில் **மற்றும்** தற்போதைய-கணு இயந்திரத்தின் ஏற்றுக்கொள் நிலை **என்றால்**

பின்னர்

திருப்புகிறது உண்மை

இல்லாவிடில்

திருப்புகிறது பொய்

இந்த வழிமுறைவரைவின் (algorithm) முக்கியப் பகுதியை விளக்கும் முன்னர் அதை ஊக்குவிக்கும் நிலைமாற்ற அட்டவணையின் (transition table) இரண்டு மாற்றங்களைக் கவனிக்கவேண்டும். முதலாவது, வெளிவரும் ε நிலைமாற்றங்கள் உள்ள கணுக்களை (nodes) உருப்படுத்தம் செய்வதற்காக, நாம் நிலைமாற்ற அட்டவணையில் புதிய ε-செங்குத்து வரிசையைச் (ε-column) சேர்க்கின்றோம். கணுவுக்கு ε-நிலைமாற்றம் (ε-transition) இருந்தால் கணுவின் கிடை வரிசைக்கு (row) ε-செங்குத்து வரிசையில் இலக்குக் கணுவைப் பட்டியல் இடுகிறோம். இரண்டாவது சேர்ப்பு, ஒரே உள்ளீட்டுக் குறியீட்டிலிருந்து (input symbol) வேறு கணுக்களுக்கு பன்முக நிலைமாற்றங்களுக்கு (multiple transitions) வேண்டி விளக்கம்

தரவேண்டும். நாம் ஒவ்வொரு சிற்றறையும் (cell) ஒரு தனி கணுவிற்குப் பதிலாக இலக்குக் கணுக்களின் (destination node) பட்டியலைக் கொண்டிருக்குமாறு செய்கின்றோம். பின்வரும் அட்டவணை படம் 9-இல் உள்ள இயந்திரத்திற்கு நிலைமாற்ற அட்டவணையைக் காட்டுகின்றது. அதற்கு ε-நிலைமாற்றங்கள் இல்லாதிருக்கையில், இயந்திர-நிலை q_2 இல் உள்ளீடு a ஆனது q_2க்கோ q_3க்கோ திரும்பக் கொண்டுசெல்வதை அது காட்டும்.

வழிமுறைவரைவு 1 ஒரு உள்ளீட்டுக் கோர்வையைப் புரிந்துகொள்ள உறுதிசெய்யப்படாத முற்றுநிலை தானியங்கியைப் பயன்படுத்துவதற்கான வழிமுறைவரைவைக் காட்டுகின்றது. உறுதிசெய்யப்படாத தெரிந்துகொள்கை என்ற செயல்பாடு செயலாக்க நடத்தையின் போது உருவாக்கப்பட்ட எல்லா தற்போதைய ஆயப்படாத விருப்பத்தேர்வுகளையும் தொடர்ந்து கண்காணிக்க மாறி நிகழ்ச்சிநிரலைப் (variable agenda) பயன்படுத்துகின்து. ஒவ்வொரு விருப்பத்தேர்வும் (தேடல் நிலை) இயந்திரத்தின் கண்ணியையும் (நிலை) நாடாவின் மீது இருப்பிடத்தையும் கொண்டிருக்கும் பட்டியலாகும். மாறி தற்போதைய-தேடல்-நிலை தற்போது ஆயப்பட்ட கிளை விருப்பத்தேர்வை உருப்படுத்தம் செய்கின்றது.

உறுதிசெய்யப்படாத தெரிந்துகொள்கை, ஒரு தொடக்கத் தேடல்-நிலையை உருவாக்கியும் நிகழ்ச்சிநிரலில் (agenda) அதை வைத்தும் தொடங்குகின்றது. இதுவரை நாம் நிகழ்ச்சிநிரலில் தேடல்-நிலைகள் எந்த ஒழுங்கில் வைக்கப்பட்டுள்ளது என்று குறிபிடவில்லை. இந்தத் தேடல்-நிலை, (search-state) எந்திரத்தின் தொடக்க எந்திர-நிலையையும் (machine-state) நாடாவின் தொடக்கத்திற்கான சுட்டியும் (pointer) கொண்டிருக்கும். 'அடுத்தது' (NEXT) என்ற செயல்பாடு நிகழ்ச்சிநிரலிலிருந்து ஒரு உருப்படியை (item) மீட்டெடுக்கவும் அதை மாறி தற்கால-தேடல்-நிலைக்கு (variable current-search-state) ஒதுக்கவும் அழைக்கப்படும்.

உறுதிசெய்யப்பட்ட தெரிந்துகொள்கையைப் (D-RECOGNIZE) பொறுத்தவரை முக்கிய கண்ணியின் முதல் செயல்பாடு நாடாவின் முழு பொருளடக்கமும் வெற்றிகரமாகப் தெரிந்துகொள்ளப்படும் போது உறுதிசெய்யவேண்டும். இது ஏற்றுகொள்-நிலை? (ACCEPT-STATE?) என்பதன் வழி அழைத்துச் செய்யப்படும்; தற்போதைய தேடல்-நிலை (current search-state), ஏற்றுக்கொள்ளும் எந்திர நிலையையும் நாடாவின் இறுதிக்குச் சுட்டிக்காட்டியையும் கொண்டிருந்தால் இது ஏற்றுக்கொள் (accept) என்பதைத் திருப்பும். நாம் செய்துமுடிக்கவில்லை என்றால் உருவாக்கு-புதிய-நிலைகள் (GENERATE-NEW-STATES) என்பதை அழைத்துச் சாத்தியமான ஒரு கண அடுத்த நடவடிக்கைகளை எந்திரம் உருவாக்கும்; இது நிலைமாற்ற அட்டவணையிலிருந்து எந்த ε-நிலைமாற்றங்களுக்கும் எந்த இயல்பான உள்ளீட்டுக் குறியீட்டு

நிலைமாற்றங்களுக்கும் தேடல்-நிலைகளை உருவாக்கும். இந்த எல்லா தேடல்-நிலைப் பட்டியல்களும் (tuples) தற்போதைய நிகழ்ச்சிநிரலில் சேர்க்கப்படும்.

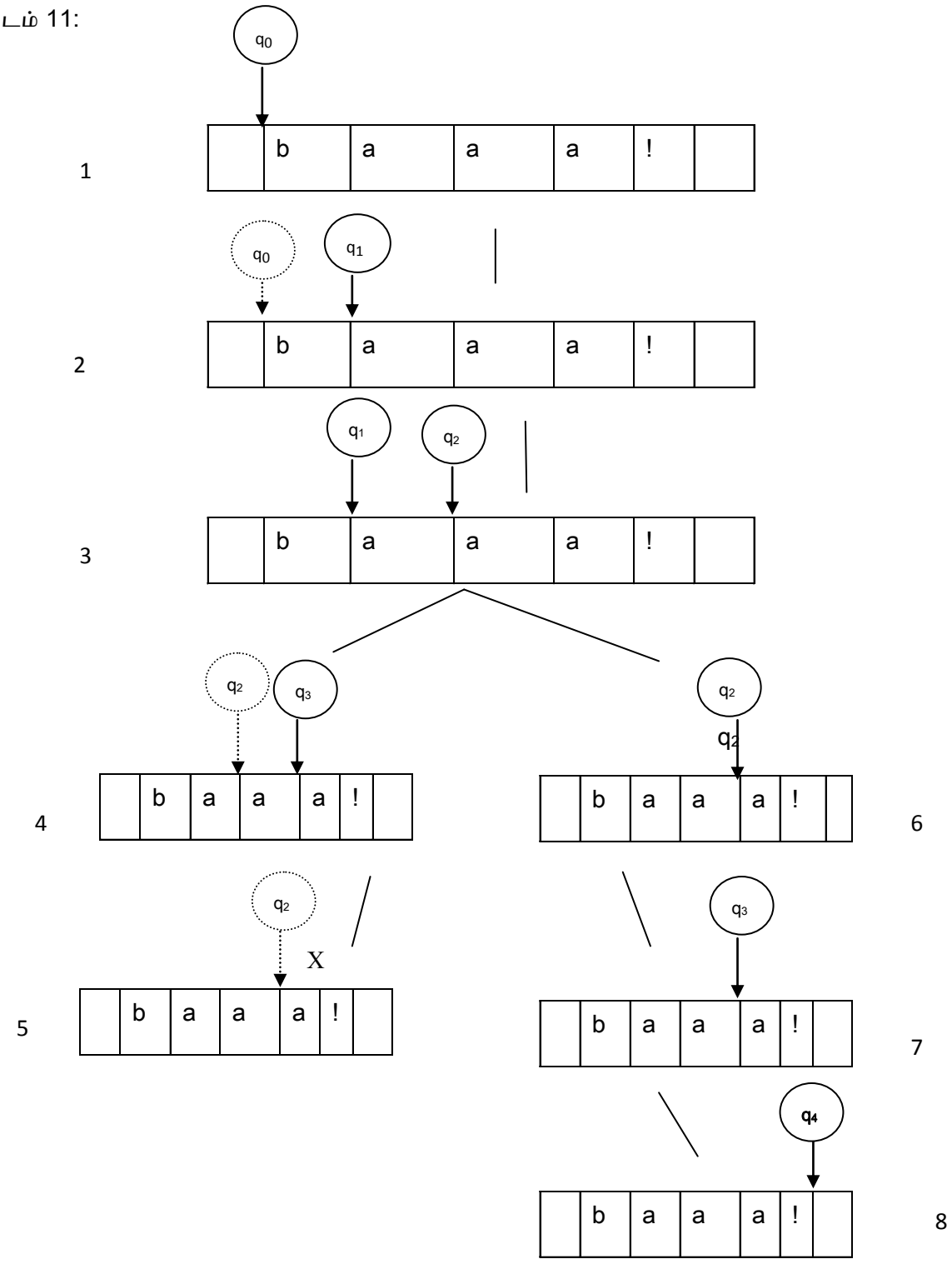
இறுதியாக, நாம் நிகழ்ச்சிநிரலிலிருந்து செயலாக்கத்திற்குப் புதிய தேடல்-நிலையைப் பெற முயற்சிக்கின்றோம். நிகழ்ச்சிநிரல் வெற்றிடம் என்றால் நமக்கு விருப்புகள் இல்லாது போகும் மற்றும் உள்ளீட்டை நிராகரிக்க வேண்டி வரும். இல்லையெனில், ஆயப்படாத விருப்பு தெரிந்தெடுக்கப்பட்டுக் கண்ணி தொடரும்.

ஏன் உறுதிசெய்யப்படாத தெரிந்துகொள்கை நிகழ்ச்சிநிரல் வெற்றிடமாகக் காணப்பட்டால் மட்டுமே நிராகரிக்கும் மதிபீட்டை திருப்புகிறது என்பதைப் புரிந்துகொள்வது முக்கியமாகும். உறுதிசெய்யப்பட்ட தெரிந்துகொள்கையைப் போல் அல்லாமல் அது ஒரு ஏற்றுகொள்ளாத எந்திர நிலையிலோ அல்லது அதால் எதாவது எந்திர நிலையிலிருந்து நாடாவை முன்னேற்ற இயலவில்லை என்றாலோ அது நிரகரி என்பதை திருப்பும். ஏனென்றால் உறுதிசெய்யப்படாத நேர்வில் இம்மாதிரியான தடைகள் தரப்பட்ட வழியில் தோல்வியைச் சுட்டிக்காடுமேயன்றி முற்றிலும் தோல்வியைச் சுட்டிக்காட்டாது. எல்லா சாத்தியமான விருப்பத்தேர்வுகளும் ஆயப்பட்டு இல்லாதிருந்தால் மட்டுமே நாம் நிராகரிப்பை உறுதி செய்ய இயலும்.

	உள்ளீடு			
நிலை	b	a	!	ε
0	1	Ø	Ø	Ø
1	Ø	2	Ø	Ø
2	Ø	2,3	Ø	Ø
3	Ø	Ø	4	Ø
4	Ø	Ø	Ø	Ø

பின்வருவது ஆட்டின் மொழியின் உள்ளீட்டைக் கையாளுவதைக் காட்டும் உறுதி செய்யப்படாத புரிந்துகொள்கையின் எடுத்துக்காட்டாகும் (Jurafsky and Martin 2000: 45).

படம் 11:


q0
b a a a !
1
q0 q1
b a a a !
2
q1 q2
b a a a !
3
q2 q3
b a a a !
4
q2
X
b a a a !
5
q2
q2
b a a a !
6
q3
b a a a !
7
q4
b a a a !
8

baaa! என்ற உள்ளீட்டைக் கையாள முயற்சிப்பதால் உறுதி செய்யப்படாத புரிந்துகொள்கையின் முன்னேற்றத்தைப் படம் 11 எடுத்துக்காட்டுகின்றது. ஒவ்வொரு நாடாவும் முன்னேற்றத்தில் தரப்பட்ட புள்ளியிடத்தில் வழிமுறையின் நிலையை எடுத்துக்காட்டுகின்றது. தற்போதைய-தேடல்-நிலை (current-search-state) மாறி நாடாவில் முன்னேற்றத்தை உருப்படுத்தம் செய்யும் அம்புக்குறியுடன் எந்திர நிலையை உருப்படுத்தம் செய்யும் கெட்டியான குமிழிகளால் தீர்மானிக்கப்படுகின்றது. படத்தில் தாழும் ஒவ்வொரு நாடாவும் தற்போதைய தேடல் நிலையிலிருந்து அடுத்ததற்கு முன்னேறுவதை உருப்படுத்தம் செய்கின்றது.

நாடாவின் மீது இரண்டாவதில் நோக்கும் போது வழிமுறை q_2ஐ கண்டுபிடிப்பது வரை கூறத்தக்க ஒன்றும் நிகழ்வதில்லை. நிலைமாற்ற அட்டவணை [q_2,a]க்கு வேண்டி பதிவைப் பரிசோதிப்பது q_2, q_3 என்ற இரண்டையும் திருப்பும். தேடல் நிலைகள் இந்த ஒவ்வொரு விருப்பதேர்வுகளுக்கு வேண்டியும் உருவாக்கப்படுகின்றது. எதிர்பாரதவிதமாக நமது வழிமுறை q_3 நிலைக்கு நகர்வதைத் தெரிவுசெய்யும்; நிலைமாற்ற அட்டவணை [q_2,a]க்கு வேண்டிய பதிவு காலியாக இருப்பதால் இந்த நகர்வு ஏற்பு நிலையிலோ புதியநிலையிலோ விளையாது. இந்த இடத்தில் வழிமுறை புதிய நிலையைத் தொடர நிகழ்ச்சிநிரலைக் கேட்கும். q_2 விலிருந்து q_2 க்கு திரும்பும் விருப்பத்தேர்வு நிகழ்ச்சிநிரலில் பரிசோதனை செய்யப்படாத விருப்பத்தேர்வு ஆகையால் இது அடுத்த a-க்கு முன்னேற்றம்செய்யப்பட்ட நாடா சுட்டியால் திருப்பப்படுகின்றது. எப்படியோ மோசமாக உறுதிசெய்யப்படாத தெரிந்துகொள்கை அதே விருப்பத்தேர்வை நேரிடுகின்றது. மாற்றமைவு அட்டவணை [q_2,a] q_2வுக்குத் திரும்புவதோ q_3வுக்கு முன்னேறுவதோ சரியான விருப்பத்தேர்வுகள் என்பதை இன்னும் சுட்டிக்காட்டுகின்றது. முன்புபோல் இரண்டையும் உருப்படுத்தம் செய்யும் நிலைகள் நிகழ்ச்சிநிரலில் வைக்கப்பட்டுள்ளது. இந்த தேடல் நிலைகள் முந்தயவை போன்றவை அல்ல; ஏனென்றால் அவற்றின் நாடா அகரவரிசையின் மதிப்புகள் முன்னேற்றப்பட்டுள்ளன. தற்போது நிகழ்ச்சிநிரல் அடுத்த நகர்வாக q_3க்கு நகர்வதைத் தருகின்றது. q_4 க்கு நகர்தலும் வெற்றியும் நாடாவாலும் நிலைமாற்ற அட்டவணையாலும் தனித்தன்மையுடன் தீர்மானிக்கப்படுகின்றது.

2.4.2.6. தேடலாகத் தெரிந்துகொள்கை (Reconition as search)

உறுதிசெய்யப்படாத தெரிந்துகொள்கை ஒரு பொறிவழியாக எல்லாச் சாத்தியமான வழிகளையும் முறையாக ஆய்வதற்கு ஒரு வழியைத் தந்து சீரான மொழியின் கோவைகளைப் புரிந்துகொள்ளும் செயலை நிறைவேற்றுகின்றது. இந்த ஆய்வு ஒரு ஏற்றுக்கொள்ளும் நிலையில் இறுதியுறும் ஒரு வழியைத் தந்தால் அது கோவையை ஏற்கும்; இலையென்றால் மறுத்து விடும். இந்த முறையான ஆய்வு நிகழ்ச்சிநிரல் உபாயத்தால் சாத்தியமாக்கப்படுகின்றது; இது ஒவ்வொரு மறு செய்கையிலும் ஒரு ஆராய்வதற்கான பகுதி பாதையைத் தேர்ந்தெடுக்கிறது, இன்னும் மீதமுள்ளவற்றைக் கண்டுபிடிக்கப்படாத பகுதி பாதைகளாகக் கண்காணிக்கும்.

தீர்வுகளைத் திட்டமிட்டு தேடுவதன் மூலம் செயல்படும் உறுதிசெய்யப்படாத தெரிந்துகொள்கை போன்ற வழிமுறைகள் நிலை-இட-தேடல் வழிமுறைகள் (state-space-search

algorithms) என அழைக்கப்படுகின்றன. அத்தகைய வழிமுறைகளில் சிக்கல் வரையறை சாத்தியமான தீர்வுகளின் இடத்தை உருவாக்குகிறது; இந்த இடத்தை ஆராய்வதே குறிக்கோள்; இடம் முழுமையாக ஆராயப்படும் போது ஒன்று உள்ளீட்டைக் கண்டறிவோஅல்லது நிராகரிக்கவோ இருக்கும்போது பதிலைத் தருகிறது. உறுதிசெய்யப்படாத தெரிந்துகொள்கையில், தேடல் நிலைகள் உள்ளீட்டு நாடாவில் உள்ள நிலைகளுடன் இயந்திர நிலைகளின் இணைப்புகளைக் கொண்டுள்ளன. கேள்விக்குரிய இயந்திரம் கொடுக்கப்பட்டால் நிலை-இடம் அனைத்து சாத்தியமான இயந்திர நிலை மற்றும் நாடா நிலைகளின் இணைகளையும் கொண்டிருக்கும். தேடலின் குறிக்கோள் நாடா நிலையின் இறுதியுடன் ஏற்றுக்கொள்ளும் நிலையின் இணைவதைத் தேடிக்கொண்டு ஒரு நிலையிலிருந்து மற்றொரு நிலைக்கு இந்த இடத்தின் வழியாகச் செல்ல வேண்டும்.

இம்மாதிரியான வழிமுறைகளின் திறன் நிலைகள் இடத்தில் நிலைகளின் கருதப்படும் ஒழுங்கு ஆகும். நிலைகளின் மோசமான ஒழுங்கு வெறிகரமான தீர்வு கண்டுபிடிக்கப்படுவதற்கு முன் பல எண்ணிக்கையிலான வெற்றிதராத நிலைகளின் பரிசோதனைக்குக் கொண்டு செல்லும். துரதிஷ்டவசமாக பொதுவாக ஒரு நல்ல விருப்பத்தேர்வை மோசமான ஒன்றிலிருந்து வேறுபடுத்திக் கூற இயலாது; பெரும்பாலும் நாம் செய்ய இயலுவது ஒவ்வொரு சாத்தியமான தீர்வுகளையும் கருதுவதுதான்.

உறுதிசெய்யப்படாத தெரிந்துகொள்கையில் நிலைகளின் ஒழுங்கு இதுவரைக் குறிப்பிடப்படவில்லை. உருவாக்கப்படுவதால் ஆயப்படாத நிலைகள் மட்டும்தான் நிகழ்ச்சிநிரலுடன் சேர்க்கப்படுகின்றது; நாம் கேட்கும் போது 'அடுத்தது' (NEXT) என்ற செயல்பாடு ஆயப்படாத நிகழ்ச்சிநிரலைத்தான் திருப்பித்தருகின்றது. எவ்வாறு 'அடுத்தது' என்ற செயல்பாடு வரையறை விளக்கம் செய்யப்படுகின்றது என்பது கேள்வி. மிக சமீபத்தில் உருவாக்கப்பட்டவைதான் அடுத்ததாகக் கருதப்படுகின்ற நிலைகள் என்ற ஒழுங்கு உபாயத்தைக் கருத்தில் கொள்ளவும். இவ்வகையிலான கொள்கையைப் புதிதாக உருவாக்கப்பட்ட நிலைகளை நிகழ்ச்சிநிரலில் முன்னிலிருக்கும் படி வைத்தும் அழைக்கும் போது நிகழ்ச்சிநிரலின் முன்பில் 'அடுத்தது' திரும்பும்படிக்கும் செயல்படுத்தலாம். இவ்வாறு நிகழ்ச்சிநிரல் ஆழம்-முதல் தேடல் அல்லது இறுதி உள் முதல் வெளி (Last In First Out (LIFO) உபாயம் என்று கூறப்படுகின்றது.

இம்மாதிரியான் உபாயம் உருவாக்கப் படுகையில் புதிதாக உருவாக்கப்படும் முன்னணிகளைப் பின்பற்றி தேடல் வெளிகளாகப் பிரிகின்றது. தற்போதைய முன்னணி வழியான முன்னேற்றம் தடைசெய்யப்படும் போது அது முந்தைய விருப்புகளைக் கருதும் படிக்குத் திரும்பும்.

படம் 11-இல் காட்டியுள்ளபடி baaa! என்ற கோர்வையின் மீது உறுதிசெய்யப்படாத தெரிந்துகொள்கையின் செயல்படுத்தலின் சுவடு ஆழம்-முதல் தேடலை எடுத்துக்காட்டுகின்றது. q_2 வில் நிற்கவேண்டுமா அல்லது q_3 வுக்கு முன்னேறவேண்டுமா என்று தீர்மானிக்க நேரும்போது ba என்பதைப் பார்த்தபின் வழிமுறை முதல் விருப்பத்தேர்வு புள்ளியிடத்தை நேரிடுகின்றது. இந்தப் புள்ளியிடத்தில் அது ஒரு மாற்றைத் தெரிவு செய்கின்றது; அது தவறு என்பதைத் தெரிந்துகொள்வது வரை அதைப் பின்பற்றுகின்றது. வழிமுறை திரும்பிச் சென்று வேறு ஒரு பழைய மாற்றை முயல்கின்றது.

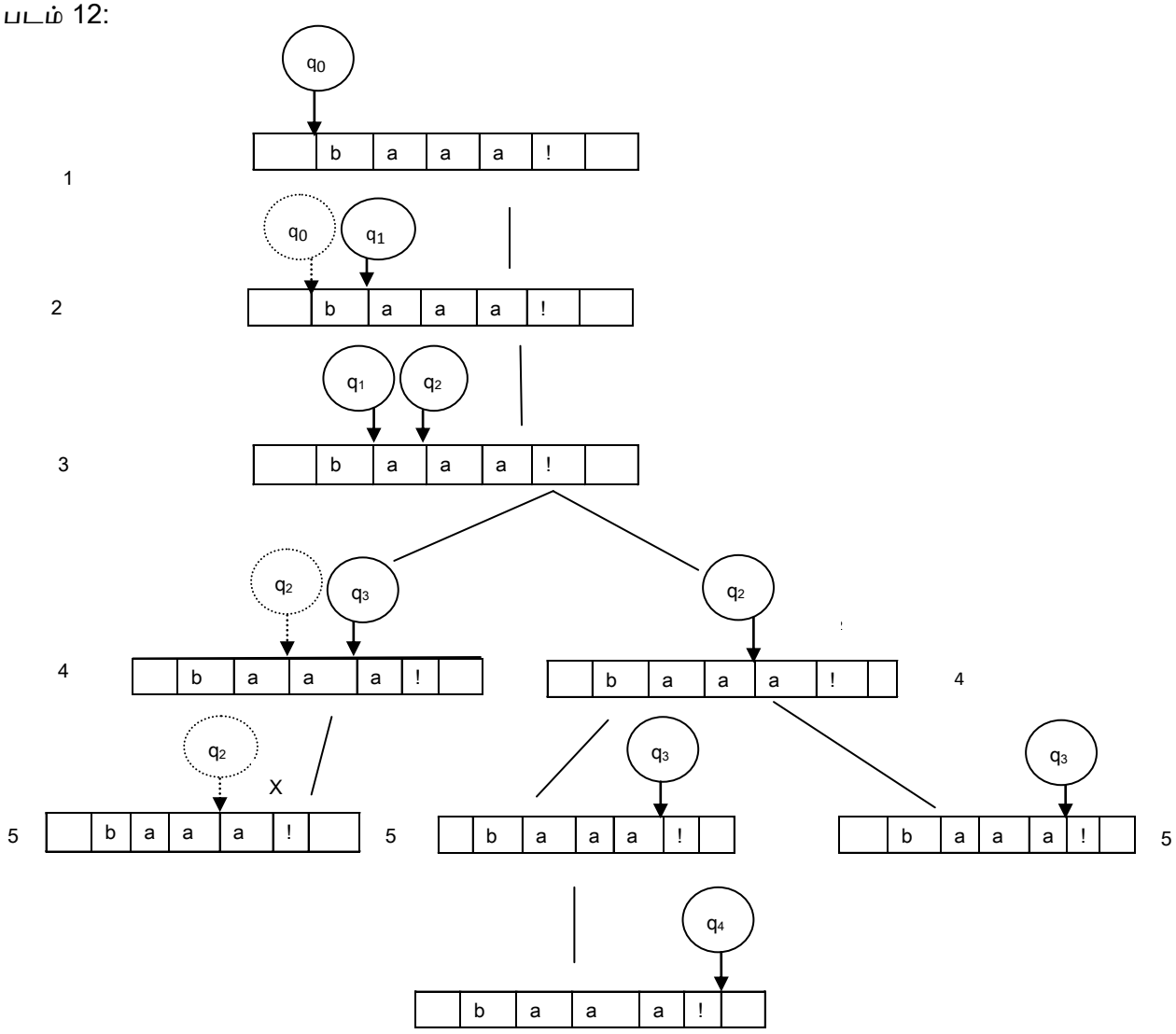
ஆழம் முதல் உபாயங்கள் ஒரு முக்கியமான ஆபத்தைக் கொண்டிருக்கின்றது: சில சூழல்களில் அவை ஒரு எல்லையற்ற கண்ணியில் நுழையக்கூடும். ஒரு தேடல் நிலை தற்செயலாக மீள்நிகழலாம் எனும்படி தேடல் வெளி/இடம் அமைக்கப்படிருந்தால் அல்லது எலையற்ற தேடல் நிலைகள் இருந்தால் இது சத்தியமாகும்.

தேடல் வெளியில்/இடத்தில் நிலைகளை ஒழுங்கு செய்யும் இரண்டாவது வழி நிலைகளை அவை உருவாக்கப்படும் ஒழுங்கில் கருதுவதாகும். நிகழ்ச்சிநிரலின் பின்பகுதியில் புதிதாக உருவாக்கப்பட்ட நிலைகளை வைத்தும் நிகழ்ச்சிநிரலின் முன்பகுதியில் 'அடுத்தது' திருப்படி செய்தும் அவ்விதமான கொள்கையை நிறைவேற்றலாம். இவ்வாறு நிகழ்ச்சிநிரல் வரிசை வழி நிறைவேற்றப்படுகின்றது. இது பொதுவாக அகலம்-முதல் தேடல் (breadth-first search) அல்லது முதல் உள் முதல் வெளி (First In First Out (FIFO) உபாயம் என்று பொதுவாகக் குறிப்பிடப்படுகின்றது. படம் 12இல் காட்டியுள்ளது போன்று baaa! என்ற கோர்வையின் மீது உறுதிசெய்யப்படாத தெரிந்துகொள்கையின் செயல்படுத்தலின் வேறு சுவட்டைக் கருத்தில் கொள்ளவும். மீண்டும் வழிமுறை ba என்பதைப் பார்த்தபின் q_2 வில் நிற்கவேண்டுமா அல்லது q_3வுக்கு முன்னேறவேண்டுமா என்று தீர்மானிக்க நேரும்போது வழிமுறை முதல் விருப்பத்தேர்வு புள்ளியிடத்தை நேரிடுகின்றது. இப்போது ஒரு விருப்பத்தேர்வைத் தேர்ந்தெடுத்து பின்பற்றுவதற்குப் பதிலாக நாம் ஒரு நேரத்தில் தேடல் கிளையமைப்பின் ஒரு பகுதியை விரித்து எல்லா சாத்தியமான விருப்பத்தேர்வுகளையும் பரிசோதிப்பதைக் கற்பனை செய்கின்றோம்.

ஆழம்-முதல் தேடல் போன்று, அகலம்-முதல் தேடலும் குறைகளைக் கொண்டுள்ளன. ஆழம்-முதல் போன்று நிலை-இடம்/வெளி (state-space) எல்லையற்றது என்றால் தேடல் இறுதியுறாது. மிக முக்கியமாக நிகழ்ச்சிநிரலின் அளவின் வளர்ச்சி காரணமாக நிலை-இடம் மிதமான அளவு பெரிது என்றாலும் தேடல் பெரிய அளவு நினைவகத்தை வேண்டும். சிறிய சிக்கல்களுக்கு ஆழம் முதல் அல்லது அகலம் முதல் தேடல் உபாயங்கள் போதுமானதாகும்;

இருப்பினும் நினைவகத்தின் பயன்பாட்டில் அதன் கூடுதல் திறமை காரணமாக பொதுவாக ஆழம் முதல் விரும்பப்படுகின்றது. பெரிய சிக்கல்களுக்கு சக்திவாய்ந்த நிரல் (programming) அல்லது A* பயன்படுத்தப்பட வேண்டும்.

படம் 12:



2.4.2.7. நிர்ணயிக்கப்பட்ட மற்றும் நிர்ணயிக்கப்படாத தானியங்கிகளை உறவுபடுத்துதல்

நிர்ணயிக்கப்படாத தானியங்கி ε-மாற்றங்கள் போன்ற நிர்ணயிக்காத அம்சங்களைக் கொண்டிருக்க அனுமதிப்பது அவற்றை நிர்ணயிக்கப்பட்ட தானியங்கிகளை விட சக்திவாய்ந்ததாக ஆக்குவது போல் தோன்றலாம். இது உண்மை இல்லை. எந்த நிர்ணயிக்கப்படாத தானியங்கிக்கும் சரியாக சமமான நிர்ணயிக்கப்பட்ட தானியங்கி உள்ளது.

உண்மையில், ஒரு நிர்ணயிக்கப்படாத தானியங்கியைச் சமமான நிர்ணயிக்கப்பட்ட தானியங்கியாக மாற்றுவதற்கான எளிய வழிமுறை உள்ளது; ஆனால் இந்த சமமான நிர்ணயிக்கப்பட்ட தானியங்கியில் உள்ள நிலைகளின் எண்ணிக்கை அதிகமாக இருக்கலாம். இயைபுகளின் நிரூபணத்திற்கு லூயிஸ் மற்றும் பாப்பாடிமிட்ரியோ (Lewis and Papadimitriou 1981) அல்லது ஹாப்கிராஃப்ட் மற்றும் உல்மேன்-ஐப் (Hopcroft and Ullman 1979) பார்க்கவும்). இருப்பினும், நிரூபணத்தின் அடிப்படை உள்ளுணர்வு எவ்வாறு நிர்ணயிக்கப்படாத தானியங்கி வழியில் அதன் உள்ளீட்டைப் பகுப்பாய்வு செய்கிறது என்பது குறிப்பிடத்தக்கதாகும். நிர்ணயிக்கப்படாத தானியங்கிகளுக்கும் நிர்ணயிக்கப்பட்ட தானியங்கிகளுக்கும் இடையிலான வேறுபாடு என்னவென்றால், ஒரு உள்ளீடு i தரப்படுகையில் ஒரு நிர்ணயிக்கப்படாத தானியங்கியில் ஒரு நிலை ஒரு உள்ளீடு q_i ஒன்றுக்கு மேற்பட்ட சாத்தியமான அடுத்த நிலைகளைக் கொண்டிருக்கலாம் (எடுத்துக்காட்டாக q_a மற்றும் q_b) என்பதை நினைவுகூரவும். படம் 2.21 இல் உள்ள வழிமுறை q_a அல்லது q_b மற்றும் தேர்ந்தெடுத்தும் தேர்வு தவறாக மாறிவிட்டால் பின்நோக்கிச்சென்றும் தேர்ந்தெடுப்பதன் மூலம் இந்த சிக்கலைக் கையாண்டது. வழிமுறையின் இணையான பதிப்பு பாதைகள் இரண்டையும் ஒரே நேரத்தில் பின்பற்றும் என்று நாம்குறிப்பிட்டோம் (q_a மற்றும் q_b நோக்கி).

2.4.3. சீரான மொழிகளும் முற்றுநிலைத் தானியங்கிகளும்

மேலே பரிந்துரைத்தபடி, முறையான வெளிப்பாடுகளால் வழக்கமாக வரையறுக்கப்பட்ட மொழிகளின் வகுப்பு முற்றுநிலைத் தானியங்கியால் (நிர்ணயிக்கப்படாத தானியங்கி அல்லது நிர்ணயிக்கப்பட்ட தானியங்கி) பண்பாக்கம் செய்ய இயலும் மொழிகளின் வகுப்பை ஒத்திருக்கும். இதன் காரணமாக, இந்த மொழிகளைச் சீரான மொழிகள் (regular languages) என்று அழைக்கிறோம். சீரான மொழிகளின் வகுப்பிற்கு முறையான வரையறையை கொடுப்பதற்காக, நாம் முந்தைய இரண்டு கருத்துருகளைத் திரும்பக் குறிப்பிட வேண்டும்: இது அனைத்து குறியீடுகளின் தொகுப்பான Σ என்ற எழுத்து, Σ இல் வழக்கமாகச் சேர்க்கப்படாத வெற்று கோர்வை ε. இத்துடன் வெற்று கணம் θவுக்கு குறிப்பு செய்கின்றோம் (இது εஇலிருந்து வேறுபட்டது). முறையான மொழிகளின் வகுப்பு (அல்லது முறையான கணங்கள்) Σ க்கு மேல் பின்னர் முறையாக பின்வருமாறு அமையும்:

1. θ ஒரு முறையான மொழியாகும்.

2,$\forall$ a $\in \Sigma \cup \varepsilon$, {a} என்பது முறையான மொழியாகும்.

3. L1, L2 என்பன முறையான மொழிகளானால், பின்னர்

(அ) L1. L2 - {xy x ∈ L1. y ∈ L2} L1, L2 என்பதன் சங்கிலி இணைப்பு

(ஆ) L1 ∪ L2, L1, L2 என்பதன் இணைப்பு அல்லது பிரிப்பு

(இ) L_1^*, L1இன் க்ளீன் மூடல்

மேலே உள்ள பண்புகளை பூர்த்தி செய்யும் மொழிகளின் தொகுப்புகள் அனைத்தும் முறையான மொழிகள் ஆகும். முறையான மொழிகள் முறையான வெளிப்பாடுகளால் பண்பாக்கம் செய்ய இயலும் முறையான மொழிகளின் தொகுப்பு என்பதால் எல்லா வழக்கமான விஷயங்களும் இருக்க வேண்டும். இந்த இயலில் அறிமுகப்படுத்தப்பட்ட வெளிப்பாடு இயக்கிகள் (expression operators) (நினைவகம் தவிர) இருக்கலாம் முறையான மொழிகளை வரையறுக்கும் மூன்று செயல்பாடுகளால் செயல்படுத்தப்படுத்த இயலும்: இணைப்புச் சங்கிலி (concatenation), பிரிப்பு/சேர்ப்பு (disjunction/union) ('|' என்றும் அழைக்கப்படும்) மற்றும் க்ளீன் மூடல் (Kleen closure). எடுத்துக்காட்டாக அனைத்து எண்ணிகளும் (counters) (*, +, {n, m}) மீள்நேர்வு (repetition) கூட்டல் க்ளீன் * என்பதன் ஒரு சிறப்பு நேர்வாகும். அனைத்து நங்கூரங்களையும்/நிறுவிகளையும் (anchors) தனிப்பட்ட சிறப்பு அடையாளங்களாக கருதலாம். சதுர அடைப்புகுறிகள் [ ] ஒரு வகையான விலகல் (அதாவது [ab] என்றால் “a அல்லது b”, அல்லது a மற்றும் b இன் விலகல்) ஆகும். எனவே எந்தவொரு முறையான வெளிப்பாடும் மூன்று பழமையான செயல்பாடுகளை மட்டுமே பயன்படுத்தக்கூடிய (ஒருவேளை பெரிய) வெளிப்பாடாக மாற்ற இயலும் என்பது உண்மைதான்.

முறையான மொழிகளும் பின்வரும் செயல்பாடுகளின் கீழ் மூடப்படுகின்றன (இங்கு Σ என்பது Σ என்ற எழுத்துக்களிலிருந்து உருவாகும் அனைத்து கோர்வைகளின் எல்லையற்ற தொகுப்பு):

குறுக்குவெட்டு (intersection): L_1உம் L_2உம் முறையான மொழிகள் என்றால் $L_1 \cap L_2$ என்பதும் முறையான மொழிதான்; இது L_1 மற்றும் L_1 இரண்டிலும் இருக்கும் கோர்வைகளின் தொகுப்பைக்/கணத்தைக் கொண்ட மொழி.

வேறுபாடு (difference): L_1உம் L_2உம் முறையான மொழிகள் என்றால், L_1 - L2 என்பதும் முறையான மொழிதான்; இது L_1இல் இருக்கிற ஆனால் L_2இல் இல்லாத கோர்வைகளின் தொகுப்பைக்/கணத்தைக் கொண்ட மொழி.

நிறைவு (complementation): L_1 ஒரு முறையான மொழி என்றால் Σ^*- L_1 என்பதும் முறையான மொழி ஆகும்; இது L_1இல் இல்லாத அனைத்துக் கோர்வைகளின் தொகுப்பு/கணம் ஆகும்.

தலைகீழ் (reversal): L_1 ஒரு முறையான மொழி என்றால் L_1^R என்பதும் முறையான மொழி ஆகும்; இது L_1 இல் உள்ள அனைத்து கோர்வைகளின் தலைகீழ் தொகுப்பைக் கொண்ட மொழி ஆகும்.

முறையான வெளிப்பாடுகள் முற்று-நிலை தானியங்கிக்கு சமமானவை என்பதற்கான ஆதாரத்தை ஹாப்கிராஃப்ட் மற்றும் உல்மேன்-இல் (Hopcroft and Ullman 1979) காணலாம்; மேலும் இது இரண்டு பகுதிகளைக் கொண்டுள்ளது: ஒவ்வொரு முறையான மொழிக்கும் ஒரு தானியங்கியை உருவாக்க முடியும் என்பதையும், மறுதலையாக ஒவ்வொரு தானியங்கிக்கும் ஒரு முறையான மொழியை உருவாக்க முடியும் என்பதையும் காட்டுகிறது. நாம் ஆதாரத்தை கொடுக்கமாட்டோம்; ஆனால் முதல் பகுதியை எவ்வாறு செய்வது என்பதைக் காட்டுவதன் மூலம் உள்ளுணர்வைக் கொடுக்கிறோம்: எந்த ஒரு முறையான வெளிப்பாட்டையும் எடுத்து அதிலிருந்து ஒரு தானியங்கியை உருவாக்கலாம். உள்ளுணர்வு தூண்டல் ஆகும்: அடிப்படை நேர்வுக்கு நாம் a என்று புலக்குறிப்பு செய்யபட்ட தொடக்க நிலையையும் ஏற்கும் இறுதி நிலையையும் உருவக்கி ஒரு தனிக் குறியீட்டின் முறையான வெளிப்பாடுகளுடன் (எ.கா. வெளிப்பாடு a) பொருந்தும் ஒரு தானியங்கியை உருவாக்குகிறோம். தூண்டக்கூடிய கட்டத்தில் ஒரு முறையான வெளிப்பாட்டின் பழமையான செயல்பாடுகளையும் (இணைத்தல், ஒருங்கிணைத்தல், மூடல்) ஒரு தானியங்கியலால் போலச்செய்ய இயலும் என்று காட்டுகிறோம்.

இணைத்தல் (concatenation): முற்றுநிலைத் தானியங்கி 1-இன் அனைத்து இறுதி நிலைகளையும் முற்றுநிலைத் தானியங்கி 2 இன் ஆரம்ப நிலையுடன் ஒரு ε-நிலைமாற்றத்தால் இணைப்பதன் மூலம் ஒன்றுக்கொன்று அடுத்ததாக வரும் இரண்டு முற்றுநிலைத் தானியங்கிகளை இணைக்கின்றோம்

மூடல் (closure): நாம் முற்றுநிலைத் தானியங்கியின் அனைத்து இறுதி நிலைகளையும் ε-நிலைமாற்றத்தால் மீண்டும் ஆரம்ப நிலைகளுடன் இணைக்கிறோம் (இது க்ளீனின் ஸ்டாரின் திரும்பநிகழும் பகுதியை செயல்படுத்துகிறது); பின்னர் ε- நிலைமாற்றத்தால் ஆரம்ப மற்றும் இறுதி நிலைகளுக்கு இடையே நேரடி இணைப்புகளைச் செய்கிறோம் (இது பூஜ்ஜிய நிகழ்வுகளைக் கொண்டிருப்பதற்கான சாத்தியத்தை செயல்படுத்துகிறது). அதற்கு பதிலாக க்ளீன்-பிளஸை செயல்படுத்த இந்தக் கடைசி பகுதியை விட்டு விடுகிறோம்.

கூட்டுறவு: நாம் ஒரு புதிய ஆரம்ப நிலை q'_0-ஐச் சேர்க்கிறோம்; பின்னர் அதிலிருந்து புதிய நிலைமாற்றங்களை சேர வேண்டிய இரண்டு இயந்திரங்களின் எல்லா முந்தைய ஆரம்ப நிலைகளுடன் சேர்க்கிறோம்.

2.4.4. சுருக்கம்

இவ்வியலில் கணினிசார் மொழியாய்வின் மிக முக்கியமான அடிப்படைக் கருத்துருவான முற்றுத் தானியங்கி அறிமுகப்படுத்தப்பட்டுள்ளது. சீரான வெளிப்பாடு மொழி அமைப்பொழுங்கு பொருத்தத்திற்கு (pattern matching) சக்தி வாய்ந்த கருவி ஆகும். சீரான வெளிப்பாடுகளில் அடிப்படையான செயல்பாடுகள் குறியீடுகளை இணைத்தல் (concatenation of symbols), குறியீடுகளை பிரித்தல் (disjunction of symbols) ([ ]), /, and .), எண்ணிகள் (counters) (*,+, { n-m}, நிறுவிகள்/நங்கூரங்கள் (anchors) ^, $) மற்றும் முன்வருகை இயக்கிகள் (precedence operators) ((,)) என் அடிப்படைச் செயல்பாடுகள்/இயக்கங்கள் சீரான வெளிப்பாட்டில் அடங்கும். எந்த சீரான வெளிப்பாட்டையும் ஒரு முற்றுநிலைத் தானியங்கியாக உருப்படுத்தம் செய்யலாம். சீரான வெளிப்பாட்டின் பாகமாகக் கருதப்படுகின்ற (\1, உடன்கூடிய ( )) நினைவகம் (memory) உயர்ந்த (முன்னேற்றமான) இயக்கமாகும்; ஆனால் அதை முற்றுநிலைத் தானியங்கியாக உருப்படுத்தம் செய்ய இயலாது. ஒரு தானியங்கி உட்படையாக (implicity) ஒரு முறையான மொழியை (formal language) தானியங்கி ஏற்றுக் கொள்கின்ற கோர்வைகளின் கணம்/குழுமம் என்று விளக்குகின்றது. ஒரு தானியங்கி எழுத்துக்கள், சொற்கள் வரைபடச்சாயல்கள் (graphic images) என்பன உள்ளடக்கிய எந்தக் குறியீடுகளின் கணங்களையும்/குழுமங்களையும் சொற்றொகுதியாகப் (vocabulary) பயன்படுத்தலாம். உறுதி செய்யப்பட்ட தானியங்கி (DFSA) அது இருக்கிற நிலையால் முழுவதுமாகத் தீர்மானிக்கப்படுகின்றது. உறுதி செய்யப்படாத தானியங்கி (NFSA) நிகழ்நிலையும் (current state) அடுத்த உள்ளீடு (Next input) தரப்படும் போது அது எடுக்க வேண்டிய பல் வழிகளில் (multiple paths) ஒன்றைத் தேர்ந்தெடுக்க வேண்டும். உறுதி செய்யப்படாத தானியங்கியையும் உறுதி செய்யப்பட்ட தானியங்கியாக மாற்ற இயலும். நிகழ்ச்சிநிரலில் (agenda) எந்த உறுதி செய்யப்படாத தானியங்கியைத் தெரிந்தெடுக்க வேண்டும் என்ற நிரக்கைப் பரிசோதிப்பது (explore) அதன் தேடல் நடவடிக்கையை/ தந்திரத்தை (search strategy) விளக்குகின்றது. ஆழம் முதல் என்ற தேடல் (depth-firsrt search) அல்லது Last-in-First-Out (LIFO)) நிகழ்ச்சிநிரல் ஒரு குவியல் (agenda–as-stack) என்பதுடன் பொருத்தமுடையது (corresponds); அகலம் முதல் என்ற தேடல் (Breadth-first-search அல்லது First-in-First-Out (FIFO)) என்ற தேடல் நிகழச்சிநிரல் ஒரு வரிசை (agenda–as–gueue) என்பதுடன் பொருத்தமுடையது. எந்த சீரான வெளிப்பாட்டையும் தானாகவே உறுதி செய்யப்படாத தானியங்கியாகவும் உறுதி செய்யப்பட்ட தானியங்கியாகவும் தொகுக்க இயலும்.

இயல் 3

கணினிசார் ஒலியனியல்

3.1. அறிமுகம்

பேச்சு சொல் பேச்சின் சிறு அலகுகளால் ஒருங்கிணைக்கப்பட்டது என்ற ஒலி அடிப்படையிலான எழுதும் ஒழுங்குமுறையில் உள்ளுறையும் கருத்து ஒலியனியலின் எல்லாத் தற்காலக் கோட்பாடுகளையும் உள்ளடக்கும் யுஆர்-கோட்பாடு ஆகும். நாம் இங்கு தற்காலப் பேச்சறிதல் மற்றும் பேச்சுருவாக்கத் தொழில்நுட்பத்தையும் கணினி ஒலியனியல் (computational phonology) என்ற தொடர்புள்ள மொழியியலின் கிளையையும் அறிந்துகொள்ள அடிப்படை உள்ளறிவையும் வழிமுறை வரைவுகளையும் அறிமுகப்படுத்தப் போகின்றோம்.

தானியங்கு பேச்சறியும் முக்கியமான செயல்பாடு ஒலியியக்க அலைவடிவை உள்ளீடாக ஏற்றுச் சொற்களின் கோர்வையை வெளியீடாக உற்பத்தி செய்யும். எதிரிடையாக உரையிலிலிருந்து பேச்சுருவாக்கம் என்ற முக்கியச் செயல்பாடு உரைச் சொற்களின் கோர்வையை உள்ளீடாக ஏற்று ஒலியியக்க அலைகளை வெளியீடாகப் பெறுவதாகும். பேச்சைப் புரிந்து கொள்ளுதல் மற்றும் உருவாக்குதல் இவற்றின் பயன்பாடு பல்வகைப்பட்டதாகும்: தானியங்கு கேட்டெழுத்துதல் (automatic dictation), ஒலிபெயர்ப்பு, கணிப்பொறிக்குப் பேச்சு அடிப்படையிலான இடைமுகங்கள் மற்றும் தொலை பேசிகள், பேச்சுக் குறையுள்ளவர்களுக்குக் குரல்சார்ந்த உள்ளீடு மற்றும் வெளியீடு என்பனவாகும்.

இங்கு நாம் பேச்சறிவான் (speech recognizer) மற்றும் உரையிலிருந்து பேச்சு (text-to-speech) ஒழுங்குமுறைகள் என்பனவற்றில் கவனம் செலுத்துவோம்: எவ்வாறு சொற்கள் ஒலிகள் (phones) என்ற தனிப்பட்ட பேச்சு அலகின் அடிப்படையில் உச்சரிக்கப்படுகிறது? பேச்சறியும் ஒழுங்குமுறை அது புரிந்துகொள்ள இயலும் ஒவ்வொரு சொல்லுக்கும் உச்சரிப்பை வேண்டும். உரையிலிருந்து பேச்சு ஒழுங்குமுறை ஒவ்வொரு சொல்லுக்கும் ஒரு உச்சரிப்பை வேண்டும். நாம் இங்கு உச்சரிப்பை விளக்க ஒலி நெடுங்கணக்குகளை (phonetic alphabets) ஒலியியல் (phonetics) களத்தின் பகுதியாக அறிமுகப்படுத்த வேண்டி வரும். மேலும் நாம் ஒலிப்புமுறை ஒலியியல் (articulatory phonetics) என்பதையும் அறிய வேண்டி வரும். இது எவ்வாறு வாயிலுள்ள ஒலிப்பான்களால் பேச்சொலிகள் உருவாக்கப்படுகிறது என்பதை ஆயும்.

ஒரு குறிப்பிட்ட ஒலி எப்பொழுதும் ஒவ்வொரு சூழலிலும் ஒரேமாதிரி உச்சரிக்கப்படுகிறது என்றிருக்குமானால் உச்சரிப்பை மாதிரிப்படுத்துவது எளிதாகும். ஆனால் இது அவ்வாறில்லை. எடுத்துக்காட்டாக, தமிழில் [த்] என்ற ஒலி வேறுப்பட்ட ஒலிச் சூழல்களில் மிக வேறுபட்டு மெய்ப்படுத்தம் செய்யப்படுகிறது. மொழியியலில் ஒலியனியல் (phonology) களம் வேறுபட்ட சூழலில் ஒலிகள் வேறுபட்டு மெய்படுத்தம் செய்யப்படுகின்றது என்பதையும் எவ்வாறு ஒலிகளின் இவ்வொழுங்குமுறை மீதி இலக்கணத்துடன் தொடர்புடையது என்பதையும் விளக்குகிறது. நாம் இவ்வேறுப்பட்ட மெய்ப்படுத்தங்களை விளக்க ஒலியனியல் விதிகளை (phonological rules) எழுத வேண்டும்.

கணினி வழி ஒலியனியல் ஆய்வு கணினி ஒலியனியல் (computational phonology) என்று அறியப்படும். கணினி ஒலியனியலின் முக்கியமான பகுதி ஒலியனியல் விதிகளை மாதிரிப்படுத்தும் கணினி இயங்குமுறைகளின் ஆய்வாகும். நாம் எவ்வாறு எழுத்துக்கூட்டல் மாற்றிகள் (spelling-rule transducers) ஒலியனியலை மாதிரிப்படுத்த இயலும் என்று காட்ட வேண்டும். பின்னர் ஒலியனியல் கற்றலின் (phonological learning) கணினி மாதிரிகளை விளக்க வேண்டும்: எவ்வாறு ஒலியனியல் விதிகள் இயந்திரக் கற்றல் வழிமுறைவரைவுகளால் தானாக ஊக்கப்படுத்த இயலும். இறுதியாக நாம் உரையிலிருந்து பேச்சு ஒழுங்குமுறைகளில் ஒரு முக்கியமான சிக்கலுக்கு வேண்டி ஒலியனியலின் மாற்றி அடிப்படையிலான மாதிரியைப் transducer-based model) பிரயோகிக்க வேண்டும்: எழுத்துக்களின் கோர்வைகளிலிருந்து ஒலிகளின் கோர்வைகளைப் பொருத்துதல். நாம் முதலில் பெரிய ஒலிப்பு அகராதியை உருவாக்குவதில் உள்ள சிக்கல்களை அளவிட வேண்டும். பின்னர் எவ்வாறு மாற்றி அடிப்படையிலான சொற்களஞ்சியங்களும் எழுத்துக்கூட்டல் அலகுகளும் எழுத்திலிருந்து உச்சரிப்புக்குப் பொருத்துவதற்கு உச்சரிப்புகளால் பெரிதுபடுத்தப்பட வேண்டும் என்று காட்ட வேண்டும்.

கணினிசார் ஒலியனியல் என்ற தலைப்பு 'ஒலியனியல்' என்ற பாட எல்லையைத்தாண்டி ஒலியியலையும் சேர்த்து இங்கு பயன்படுத்தப்பட்டுள்ளது. பேச்சாய்வு ஒலிப்பியலுடனும் ஒலியியக்கவியலுடன் தொடர்புள்ளது என்ற காரணத்தால் ஒலிப்பியல் மற்றும் ஒலியியக்கவியல் குறித்த சுருக்கம் இங்கு தரப்பட்டுள்ளது.

3.2. ஒலியியல் ஆய்வு

மொழிகளின் ஒலிகளைப் பிறப்பித்தல், காற்றில் பரவுதல், காதால் கேட்டு உணர்தல் என்ற அடிப்படையில் வகைப்படுத்தி ஆய்வது ஒலியியல் (phonetics) எனப்படும். ஒலியியல் என்றால் ஒலிகளைப் பற்றி படிப்பது என்று பொதுவாகக் கூறலாம். மொழியியலில் பேச்சொலிகள் மட்டும் ஆயப்படுகின்றன என்பதால் ஒலியியல் என்பதைப் பேச்சொலிகளைப் பற்றி ஆய்தல் அல்லது படித்தல் என விளக்கம் தரலாம். கருத்துப்பரிமாற்ற மாதிரியில் தகவலை அனுப்புவர் மூளையில் எண்ணும் தகவலானது நரம்பணுக்களின் உதவியால் இயக்கப்படும் ஒலியுறுப்புகளால் ஒலிகளாகக் குறியாக்கம் செய்யப்பட்டு, காற்று ஊடகத்தின் வழியாகச் சென்று பெறுபவர் காதுகளை அடைந்து அவரது மூளையால் குறியத் திறவு செய்யயப்பட்டு தகவல் புரிந்துகொள்ளப்படுகின்றது. இதன் அடிப்படையில் ஒலியியலை மூன்று வகைகளாகப் பிரிக்கலாம்.

மூவகை ஒலியியல் (three types of phonetics or branches of phonetics): ஒலிகளை அதன் பிறப்பு, காற்றில் பரவுதல், காதால் உணர்தல் அடிப்படையில் மூன்றாகப் பகுக்கலாம்:

- ஒலிப்பியல் (articulatory phonetics)
- ஒலியியக்கவியல் (acoustic phonetics)
- ஒலியுணர்வியல் (auditory phonetics)

ஒலிப்பியல் (articulatory phonetics)

ஒலிகளை அவற்றின் பிறப்பு அல்லது உச்சரிப்பு அடிப்படையில் வகைப்படுத்தி ஆய்தல் ஒலிப்பியல் எனப்படும். இதனால் இது உச்சரிப்பு ஒலியியல் என்றும் ஒலி பிறப்பியல் என்றும் அறியப்படும். இவ்வாய்வுமுறை பேச்சுறுப்புக்களின் அமைப்பையும் தொழிலையும் ஆதாரமாகக் கொண்டு உடற்கூற்றியல் நோக்கில் ஒலிகளின் பிறப்பை ஆராயும். ஒலிகள் அவற்றின் ஒலிப்பிடம், ஒலிப்புமுறை அடிப்படையில் வகைப்படுத்தப்படும். ஒலிப்பியலைக் காலத்தால் முந்திய ஒலியாய்வு முறை என்பர். அனேகமாக எல்லா மொழிகளிலுள்ள மரபிலக்கணங்களிலும் ஒலிகள் இவ்வாய்வு முறையில்தான் ஆராயப்படுகின்றன. இந்திய மொழிகளில் மிகப் பழைய மொழிகளாகக் கருதப்படுகின்ற சமஸ்கிருதத்திலும் தமிழிலும் ஒலிப்பியல் முறையில் ஒலிகள் ஆராயப்படுகின்றன. இன்றைய ஒலிப்பியல் வளர்ச்சிக்கு ஆதாரமாகவும் அடிப்படையாகவும் அமைந்தது பாணினியின் ஒலிப்பியல் விளக்கமே என்பர். மரபிலக்கணங்களில் எழுத்துகளுக்கும் ஒலிகளுக்கும் இடையேயும் எழுத்துக்களின் எண்ணிக்கைக்கும் ஒலிகளின் எண்ணிக்கைக்கும் இடையேயும் நிலவி வந்த மயக்கநிலை இவ்வாய்வு முறையால் தெளிந்தது. மனிதனுடைய வாயறையில் உள்ள பல், நாக்கு, அண்ணம் போன்ற பல்வேறு இடங்களில் இருந்து பல்வேறு பேச்சொலிகள் பிறக்கின்றன. ஒரு ஒலி பிறக்கும் இடங்களையும் வகைகளையும் விளக்கி ஆராயும் ஒலியியல் பிரிவு ஒலிப்பியல் எனப்படும். ஒரு மொழியின் ஒலிப்பிடம், ஒலிப்பு முறை ஆகியவற்றைப் பற்றி விளக்குவது ஒலிப்பியல் அல்லது பேச்சொலியியல் எனப்படும். பேச்சொலிகளை உயிரொலிகள் (vowels), மெய்யொலிகள் (consonants) எனப்பிரிப்பர். தமிழ் மரபிலக்கணம் தமிழுக்கு அ, ஆ, இ, ஈ போன்ற 12 உயிரெழுத்துக்கள் இருப்பதாயும் 18 மெய்யெழுத்துக்கள் இருப்பதாயும் கூறும். ஒலிகளை முதலாவதாக அவற்றின் ஒலிப்பு முறை அடிப்படையில் உயிரொலிகள் மற்றும் மெய்யொலிகள் எனப் பகுக்கலாம்.

ஒலியியக்கவியல் (acoustic phonetics)

காற்றில் பரவும் ஒலிகளின் இயக்கத்தைப் பற்றி ஆயும் படிப்பு ஒலியக்கவியல் ஆகும். இது ஒலிகளின் ஒலியியக்க அமைப்பை ஆராய்கிறது. ஓர் ஒலியை மற்றொரு ஒலியிலிருந்து வேறாக நாம் கேட்டுணர்வதற்கு முக்கிய காரணமாக அமைவது அதன் ஒலியியக்கக் கூறுகளாகும். ஒலிகளின் பல்வேறு ஒலியியக்கக் கூறுகளை ஆராய ஒலிவர்ணனை வரைபடக்

கருவி/ஒலிநிறமாலை வரைவி (spectrograph) என்ற கருவியை ஒலியியல் அறிஞர்கள் கையாளுவர். பேச்சொலிகளின் பண்பு நலன்களை ஒலியாராய்ச்சி கருவிகளின் மூலம் கண்டறிவதே ஒலியியக்கவியலின் செயலாகும்.

ஒலியுணர்வியல் (auditory phonetics)

ஒலிகளை காதால் கேட்டு உணர்வதை ஆயும் படிப்பு ஒலியுணர்வியல் எனப்படும். இதைக் கேட்பொலியல், செவிப்புலவியல் என்றும் கூறுவர். ஒலியுணர்வியல் என்பது ஒரு ஒலி செவியால் உணரப்படும் தன்மையை விவரித்தல் ஆகும். ஒலியாராய்ச்சியில் மூன்றாவது ஆய்வு முறையாக எண்ணப்படுவது ஒலியுணர்வியலாகும். இவ்வாய்வு முறையில் ஒலிகள் அவற்றை ஒலியாராய்ச்சியாளன் கேட்கும் நிலையையும் அறிந்துணரும் வகையையும் ஆதாரமாகக் கொண்டு ஆராயப்படுகின்றன. ஒலியாராய்ச்சியில் ஒலிப்பியல் வரவேற்கப்பட்ட அளவிற்கு ஒலியுணர்வியல் வரவேற்கப்படவில்லை. பெரும்பாலான மொழி ஆராய்ச்சிகளில் ஒலிப்பியலையோ ஒலியியக்கவியலையோ அடிப்படையாகக் கொண்ட ஒலியாராய்ச்சியையே பார்க்க முடிகிறது.

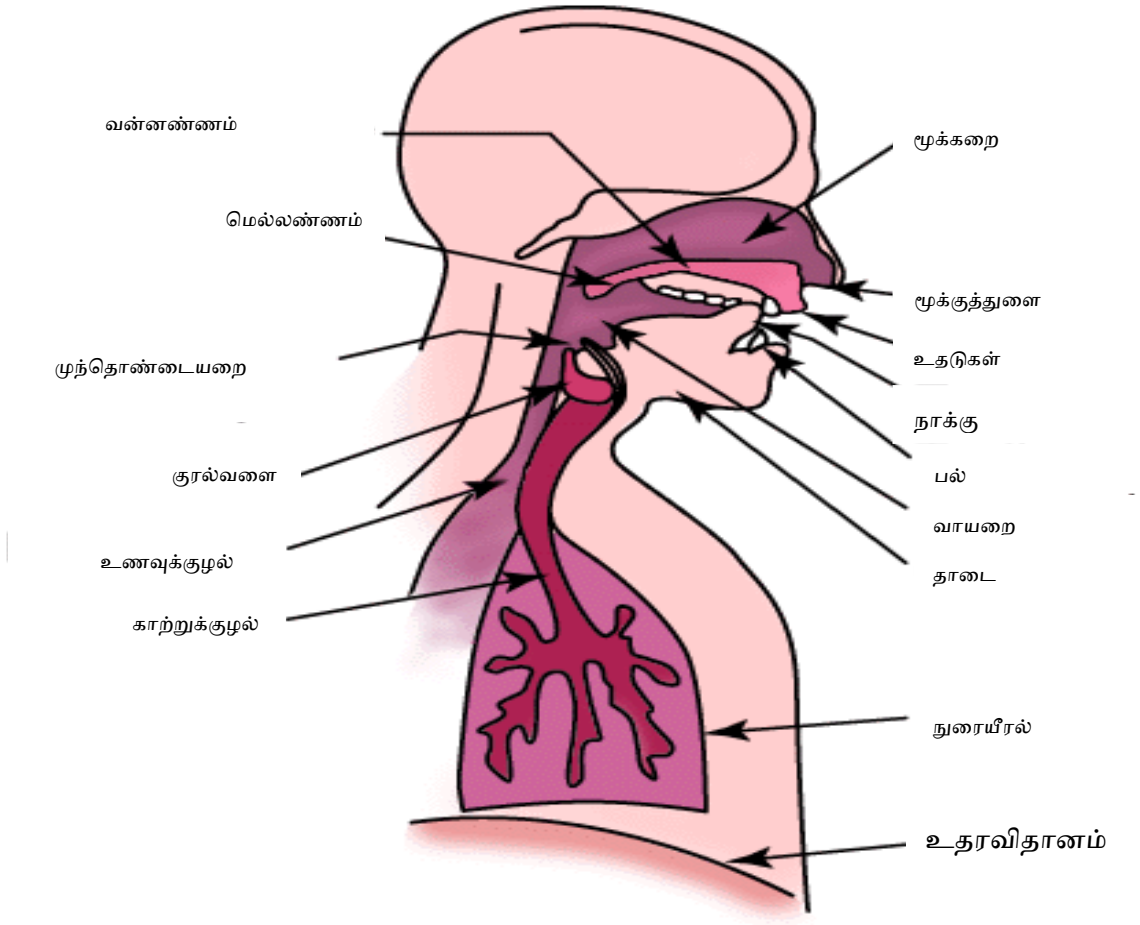
3.3. ஒலிப்பியல் ஆய்வு

ஒலிகளின் ஒலியியக்கவியல் ஆய்வு தான் நமது தலையாய நோக்கம் என்றாலும் தமிழ் ஒலிகளின் ஒலிப்பியல் ஆய்வின் பின்னணி இல்லாமல் தமிழ் ஒலிகளின் ஒலியியக்கப் பண்புக்கூறுகள் பற்றி புரிந்துகொள்வது கடினம். எனவேதான் தமிழ் ஒலிகளின் ஒலிப்பியல் ஆய்வு பற்றியும் இங்கு விளக்கப்பட்டுள்ளது. மட்டுமன்றி தமிழ் உரையிலிருந்து பேச்சு இணைப்பாக்கத்திற்கும் தமிழ் ஒலிகளின் ஒலிப்பியல் குறித்த தகவல்கள் தேவைப்படும். குறிப்பாகச் சூழல் அல்லது இலக்கண அடிப்படையில் தமிழ் ஒலிகள் மாறுவது பற்றிய விதிகள் உரையிலிருந்து பேச்சு இணைப்பாக்கத்திற்கு மிகத் தேவையாகும். இத்தகைய தேவையைத் தமிழ் ஒலிகளின் ஒலிப்பியல் ஆய்வு கருத்திற் கொண்டு விளக்கம் தருகின்றது.

முன்னர் கூறியபடி பேச்சொலிகளின் பிறப்பு பற்றி ஆய்வது ஒலிப்பியல் ஆய்வாகும். இதில் ஒலியுறுப்புகள் பற்றியும் அவை எவ்வாறு வேறுபட்ட ஒலிகளை உருவாக்குகின்றன என்பது பற்றியும் அவற்றை ஒலிப்பிடம் அடிப்படையிலும் ஒலிப்புமுறை அடிப்படையிலும் எவ்வாறு வேறுபடுத்தலாம் என்பது பற்றியும் இங்கு விளக்கப்படும். குறிப்பாகத் தமிழ் ஒலிகளின் பிறப்பியல் பற்றி இங்கு விரிவாகப் பேசப்படும். ஒலிப்பியல் ஆய்வு ஒலியியக்கவியல் ஆய்வுக்கும் உரையிலிருந்து பேச்சாய்வுக்கும் பேச்சிலிருந்து உரையாய்விற்கும் அடிப்படையாய் அமையும்.

3.3.1. ஒலியுறுப்புகள் (organs of speech)

பேச்சொலிகளை ஒலிப்பதற்குப் பயன்படும் உடல் உறுப்புகளே ஒலியுறுப்புகள் எனப்படுகின்றன. மொழியில் உண்டாக்கப் படுகின்ற ஒலிகள் பேச்சொலிகள் எனப்படும். இவ்வொலிகளை ஒலிப்பதற்கு பயன்படும் உடல் உறுப்புகள் பேச்சு உறுப்புகள் எனப்படும். இதழ், நுரையீரல், தொண்டை, வாய், மூக்கு, பல், நாக்கு, அண்ணம் போன்றவை இதில் அடங்கும். இவற்றிற்குப் பிற செயல்கள் முதன்மைச் செயல்களாக இருந்தாலும் பேச்சொலிகளை உருவாக்குவதில் இவை பெரும் பங்கு வகிக்கின்றன. இவ்வுறுப்புகளை மூன்று பிரிவுகளாக பிரிக்கலாம். 1. நெஞ்சுப் பகுதி 2. தொண்டைப் பகுதி 3. தலைப் பகுதி. இந்தப் பகுதிகளில் முறையே உயிர்த்தல் (respiration), குரல் எழுப்புதல் (voicing), ஒலித்தல் (articulation) என்ற மூன்று செயல்கள் நடைபெறுகின்றன. நுரையீரல் (lungs) முதல் உதடுகள் (lips) வரை உள்ள உறுப்புகளின் இயக்கங்களே பேச்சொலிகளைப் பிறப்பிக்கின்றன. ஒலியுறுப்புகள் என்னும் பெயர் நுரையீரல் முதல் உதடுகள் வரையுள்ள உறுப்புகளையும் அவற்றை இயக்கும் தசைகளையும் குறிக்கின்றது. இவ்வுடற் பகுதியைப் பேச்சுப்பாதை (speech tract) என்று அழைப்பர். ஒலியுறுப்புகளை எல்லாம் கீழ்வரும் பேச்சுப்பாதையின் படம் காட்டும்: படம் 4



பேச்சு உறுப்புகள் பின்வருவனவற்றை உள்ளடக்கும்: 1. இதழ் (உதடு) (lips), 2. பல் (teeth), 3. நுனியண்ணம் (alveolar ridge/alveolum), 4. இடையண்ணம் (dome), 5. கடையண்ணம் (velum), 6. உள்நாக்கு (uvula), 7. நுனிநா (tip of the tongue), 8.இடைநா (blade of the tongue), 9. கடைநா (dorsum), 10. அடிநா (root of the tongue), 11. மூக்கறை (nasal cavity), 12. மூக்கறை வாயில் (velic), 13. முன்தொண்டை (pharnx), 14.காற்றுக்குழல் மூடி (epiglottis), 15. குரல்வளை மடல்கள் (vocal chords), 16. காற்றுக் குழல் (trachea), 17. குரல் வளை (larynx/adams apple).

மூக்கு (Nose)

சுவாசிப்பதற்கும் வாசனையை அறிவதற்கும் பயன்படுகிற ஒரு உடல் உறுப்பு மூக்கு எனப்படும். ஒரு வாயறை ஒலியை எழுப்பும் போது பின்னண்ணத்தைச் சற்றுத் தாழ்த்தி மூக்கறை வழியாகவும் காற்றை வெளியிட்டு வாயறை ஒலியை ஓரளவு மூக்கொலியாக மாற்றி உச்சரித்தல் நிகழும். *அவன்* என்ற சொல்லைப் பேச்சுத் தமிழில் *அவ(ன்)* என *வ* என்ற ஒலியை மூக்கொலி இணைத்து உச்சரித்தல் எடுத்துக்காட்டு ஆகும்.

நா (tongue)

மொழியில் காணப்படுகின்ற பல ஒலிகளை உச்சரிப்பதற்கு பயன்படும் ஓர் உறுப்பு நா. பேச்சுறுப்புகளில் அண்ணத்தை நோக்கி வளையவும் குறுகவும் செய்வது நா வளை எனப்படும். குரல்வளை மடல்களின் அதிர்வோடு ஆனால் ஒலிபாதையில் எவ்விதத் தடுப்பும் உரசலும் இன்றித் தோன்றுவது உயிர்கள் எனப்படும். நாவானது நுனிநா (tip of the tongue), இடைநா (blade of the tongue), கடைநா (dorsum), அடிநா (root of the tongue) என நான்காகப் பகுக்கப்படும்.

விளிம்பு (Blade)

நுனி நாவின் முனையைச் சுற்றியும், அதனையொட்டி பின்னால் தொடர்ந்தும் அமைந்துள்ள நாவின் ஓரம் விளிம்பு ஆகும்.

இதழ் (lip)

பேசுவதற்குப் பயன்படுகின்ற உறுப்புகளில் இதழ் முக்கியமான உறுப்பாகக் கருதப்படுகிறது. மொழியில் உள்ள ஒலிகளான உயிரொலிகளையும் மெய்யொலிகளையும் உச்சரிக்கும் போது இதழின் பங்கு முக்கியமானதாக உள்ளது. பேச்சுறுப்புகளில் இதழ் அசையும் உறுப்பாகும் (active articulator). இதனை மேலிதழ் (upper lip), கீழிதழ் (lower lip) என்று இரண்டாக வகைப்படுத்தலாம். சில ஒலிகளை உச்சரிப்பதில் கீழிதழும் சில ஒலிகளை உச்சரிப்பதில் மேழிதழும் செயல்படுவதால் ஒலி வகைகளின் அடிப்படையில் இவ்வேற்றுமை முக்கியமானது எனலாம். கீழிதழ் போல் மேலிதழ் எந்த ஒலியையும் ஒலிப்பதில் தனியாகச்

செயற்படுவதில்லை. உயிரொலிகளைப் பொறுத்தவரையில் இதழ்களின் பங்கு மிகவும் முக்கியமானது. சில உயிரொலிகள் உச்சரிக்கப்படும் போது இதழ்கள் குவிந்தும் வேறு சில உயிரொலிகள் உச்சரிக்கப்படும் போது விரிந்தும் செயற்படுகின்றன.

பல் (tooth)

பேச்சுறுப்புக்களுள் மற்றொரு முக்கியமான உறுப்பு பல் ஆகும். பல் ஒரு நிலையொலிப்பி அல்லது அசையா உறுப்பு (passive articulator) ஆகும். பல்லொலிகளின் பிறப்பிடம் பல்லாகும். பல் வாயில் சொற்களை உச்சரிக்க ஏற்ற வகையில் இரு தாடைகளிலும் வரிசையாக அமைந்திருக்கும் தட்டையான அல்லது கூரிய முனை கொண்ட உறுதியான வெண்ணிற உறுப்பு ஆகும். ஒலியியலில் பல் என்பது மேற்பல் வரிசையையும் கீழ்ப்பல் வரிசையையும் குறிக்கிறது. இது அண்பல்லையும் (alveolar) அண்ணத்தையும் (palate) சுற்றி அமைந்திருக்கிறது. சில ஒலிகளின் பிறப்பாக்கத்தின் போது மேற்பல் வரிசையும் கீழ்பல் வரிசையும் ஒலிப்பிடங்களாக அமைகின்றன. இருந்தாலும் மேற்பல் வரிசையே பல ஒலிகள் ஒலிப்பதற்கு ஒலிப்பிடமாக அமைகிறது எனலாம்.

அண்பல் (alveolar)

அண்பல் பேச்சுறுப்புக்களில் ஓர் அசையா உறுப்பு (passive articulator) ஆகும். இது வாயில் மேற்பல் வரிசையை அடுத்திருக்கிறது. நுனி நா அண்பல்லைப் பொருந்த பல ஒலிகள் பிறக்கின்றன. இவ்வொலிகளை அண்பல் ஒலிகள் எனலாம்.

அண்ணம் (palate)

அண்ணம் மேற்பல் வரிசையின் அடியிலிருந்து உள்நாக்கு வரை பரந்திருக்கும் கடினமானப் பகுதி. மேற்பல் வரிசையை அடுத்திருக்கும் பகுதியை அண்பல்/ நுனியண்ணம் (alveolar) என்றும் அண்பல்லை அடுத்திருக்கும் மேற்பரப்பை இடையண்ணம் (hard palate) என்றும் இடையண்ணத்தை அடுத்திருக்கும் பகுதியை கடையண்ணம் (soft palate) என்றும் அண்ணத்தைப் பிரிக்கலாம். இவற்றுள் அண்பல்லிற்கும் இடையண்ணத்திற்கும் இல்லாத ஒரு சிறப்பு கடையண்ணத்திற்கு உண்டு. மேலும் அண்ணத்தின் மேற்பரப்பை முதல் நா அண்ணம் என்றும் நுனிநா அண்ணம் என்றும் கடை நா அண்ணம் என்றும் பிரிக்கலாம். இவை ஒலிப்பின் போது இயங்காமல் நிலையாக இருப்பதால் இவை நிலையொலிப்பி (passive articulator) எனப்படும்.

உள்நா (uvula)

உள்நா வாயறையையும் (vocal cavity) மூக்கறையையும் (nasal cavity) பிரிக்க உதவும். வாயொலிகளின் (vocal sounds) உச்சரிப்பின் போது இது மூக்கறை வாயிலை அடைத்து காற்று மூக்கறைக்குச் செல்வதை தடைசெய்யும். மூக்கொலிகளின் உச்சரிப்பின் போது இது மூக்கறை வாயிலைத் திறந்து காற்று மூக்கறை வழியாகச் செல்வதை அனுமதிக்கும். குரல்வளை மடல்கள் (vocal chords) தொண்டை குருத்தெலும்புகளால் ஆன பெட்டியாகும். இப்பெட்டியில் ஒட்டிக்கொண்டிருக்கும் இரண்டு மெல்லிய சவ்வுகளாலான பேச்சுறுப்பை குரல்வளை மடல்கள் என்பர். குரல்வளை மடல்கள் அடைப்பு நிலையில் இருக்கும்போது பிறக்கும் ஒலியை குரல்வளை அடைப்பொலி (glottal stop) என்பர். குரல்வளை மடல்கள் அகன்று விரிந்திருக்கையில் வெளியேறும் மூச்சு அதிர்வின்றி வெளியேறும். இவ்வொலி ஓசை குன்றி இருக்கும். இவ்வாறு உச்சரிக்கப்படும் ஒலிகள் குரலிலா ஒலிகள் (voiceless sounds) எனப்படும். குரல்வளை மடல்கள் அண்மைப்பட்டு இருக்கையில் வெளியேறும் காற்று அதிர்வைப்பெறுவதால் அப்போது உச்சரிக்கப்படும் ஒலி ஓசையுடனிருக்கும். அது குரலொலி (voiced sounds) எனப்படும்.

குரல் வளை மூடி (Epiglottis)

குரல் வளை மூடி பேச்சுறுப்புகளில் ஒன்று. இது மேற் தொண்டையில் அமைந்துள்ளது. சுவாசிக்கும் குரல் வளை மூடி திறந்தே இருக்கும். இதனால் மூச்சு இதன் வழியே தங்குதடையின்றி சென்று வரும்.

வாயறை (Mouth Cavity)

வாய் பேசுவதற்கான ஒரு உறுப்பு. நுரையீரலிலிருந்து வெளியேறும் மூச்சை பல்வேறு ஒலியலைகளாக மாற்றும் பல பேச்சுறுப்புகள் அடங்கிய பகுதி வாயறையாகும். வாயறையில் உருவாகும் ஒலிகளை எல்லாம் வாயொலிகள் எனலாம். உயிரொலிகளும் பெரும்பாலான மெய்யொலிகளும் வாயொலிகளாகும். உயிரொலிகள் பிறக்கும் போது நுரையீரலிலிருந்து வெளிவரும் மூச்சு வாயில் எவ்வித தடையுமின்றி வெளியேறும். மாறாக மெய்யொலிகள் பிறக்கும் போது மூச்சு வாயின் பல இடங்களில் தடைப்பட்டும் மாற்றமடைந்தும் வெளியேறும். இவ்வாறு மூச்சு வாயில் தடைப்பட்டும் தடைபடாமலும் வரும் இயல்பை அடிப்படையாக கொண்டே ஒலிகளை எல்லாம் இவ்விரு வகையாக வகைப்படுத்தலாம்.

மூக்கறை (nasal cavity)

மூக்கு துவாரங்களுக்கும் உள்நாவுக்கும் இடைப்பட்ட பகுதி மூக்கறையாகும். உள்நா வாயறையையும் (vocal cavity) மூக்கறையையும் (nasal cavity) பிரிக்க உதவும். வாயொலிகளின் (vocal sounds) உச்சரிப்பின் போது இது மூக்கறை வாயிலை அடைத்து காற்று மூக்கறைக்குச்

செல்வதைத் தடைசெய்யும். மூக்கொலிகளின் உச்சரிப்பின் போது இது மூக்கறை வாயிலைத் திறந்து காற்று மூக்கறை வழியாகச் செல்வதை அனுமதிக்கும். மூக்கறைவழியாக காற்று செல்லுகையில் உச்சரிக்கப்படும் ஒலிகள் மூக்கொலிகள் (nasal sounds) எனப்படும்.

குரல்வளைக் காற்றறை (Glottalic Chamber)

தொண்டையில் இரண்டு மெல்லிய சவ்வுகளாலான பேச்சுறுப்பைக் குரல்வளை எனலாம். நுரையீரலிலலிருந்து வரும் காற்று குரல்வளை மடல்களின் இடைவெளியாகச் செல்லும். இவ்விடைவெளியைக் காற்றறை என்று கூறலாம். குரல்வளை மடல்கள் அகலமாக விரிந்திருக்கும் போது நுரையீரலிலிருந்து வெளிவரும் மூச்சுக் காற்று தங்குதடையின்றி வெளியேறும். அப்போது மென்மையான சவ்வுகளாலான இக்குரல்வளை மடல்களில் எந்த விதமான அதிர்வும் இருப்பதில்லை. இதுவே குரல்வளைக் காற்றறை எனப்படும்.

3.3.2. மூச்சோட்ட இயக்கம் (air-stream process)

மூச்சுக்காற்றானது மூக்கு மற்றும் வாய் வழியாக விரைந்து நுரையீரலுக்குச் சென்று வருதலை மூச்சோட்டம் எனலாம். நுரையீரல் முதல் குரல்வளையுள்ள உறுப்புகளின் இயக்கத்தை மூச்சோட்ட இயக்கம் (air-stram process) என்றும் குரல்வளையில் நிகழ்வதை ஒலிப்புடமை இயக்கம் (phonation process) என்றும் குரல்வளை முதல் உதடுவரை நிகழ்வதை ஒலிப்பு இயக்கம் (articulatry process) என்றும் மூக்கறையில் நிகழ்வதை வாய்மூக்கு இயக்கம் (oral-nasla proeess) என்றும் அழைக்கப்படுகின்றன. ஒலிப்புகள் மூச்சோட்ட இயக்கத்தை மாற்றி பலவித ஒலிகளை எழுப்புகின்றது எனலாம்.

நுரையீரல் காற்று (Pulnmonic air)

மூச்சுக் காற்று நுரையீரலிலிருந்து வெளியேறுகின்ற நிலையையே நுரையீரல் காற்று எனலாம். நுரையீரல் மனித உடல் உறுப்புகளில் ஒன்று. மனிதனின் மார்புக் கூட்டினுள்ளே சவ்விலான பைகளைப் போன்ற அமைப்போடு கூடிய இரு சுவாச உறுப்புக்களை நுரையீரல்கள் என்று கூறலாம். நுரையீரல்கள் பெரும்பாலான ஒலிகள் பிறக்கும் போது மூச்சை வெளியே தள்ளுகிற செயலை மிகவும் சுறுசுறுப்போடு செய்கின்றன. மூச்சுக் குழல் வழியே இம்மூச்சு மேல் நோக்கி வந்து வாயின் பல்வேறு இடங்களில் தடைபட்டும் தடைபடாதும் பல ஒலிகளாகப் பிறக்கின்றன.

அகவுயிர்புக் காற்று (Ingressive)

சுவாசிக்கும் போது வெளியேயுள்ள காற்று வாய் அல்லது மூக்கு வழியாகத் தொண்டையை வந்தடைந்து பின்னர் நுரையீரலை வந்தடைகிறது. இவ்வாறு உட்செல்லும் காற்றை

அகவுயிர்ப்புக் காற்று எனலாம். பொதுவாக காற்று நுரையீரல்களுக்குள்ளே செல்கின்ற நிலையில் அதிகமான ஒலிகள் பிறப்பதில்லை. மிக அபூர்வமாக சில ஒலிகள் ஒலிக்கப்படலாம். ஆங்கிலத்தில் *yeah* என்ற சொல்லை காற்று நுரையீரல்களுக்குள்ளே செல்லும் நிலையில் சில ஆங்கிலேயர்கள் உச்சரிப்பதை இதற்கு எடுத்துக்காட்டாகக் கூறலாம்.

புறவுயிர்ப்புக் காற்று (Egressives)

காற்று சுவாசிக்கும்போது அகவுயிர்பினால் நுரையீரலை வந்தடைந்து காற்று மீண்டும் தொண்டைடைய அடைந்து வாய் அல்லது மூக்கு வழியே வெளியேறுகிறது. இவ்வாறு வெளியேறும் காற்றைப் புறவுயிர்ப்புக் காற்று எனலாம். பொதுவாக எல்லா ஒலிகளும் காற்று நுரையீரலிலிருந்து வெளியேறும் நிலையிலையே பிறக்கின்றன. நுரையீரலிலிருந்து வெளியேறும் காற்றின் துணையின்றியும் சில ஒலிகளை ஒலிக்க முடியும். எடுத்துக்காட்டாக tut-tut, tsk-tsk போன்ற சொட்டையொலிகளை ஒலிக்கும்போது இக்காற்றின் தேவையின்மையை உணர முடியும்.

3.3.3. ஒலிப்பிகள் (Articulators)

பேச்சொலி எழுப்பும் உறுப்புகள் ஒலிப்பி எனப்படும். எடுத்துக்காட்டாக, நாக்கு, அண்ணம், பல், உதடு போன்றவை. ஒலிப்பிகள் நிலை ஒலிப்பிகள் (fixed articulator), இயங்கு ஒலிப்பிகள் (active articulator) என இருவகைப்படும். ஒலிகள் ஒலிக்கும் போது இயங்காதிருப்பன நிலை ஒலிப்பிகளாகும். எடுத்துக்காட்டாக அண்ணம் நிலை ஒலிப்பியாகும். ஒலிகள் ஒலிக்கும் போது இயங்குவன இயக்க ஒலிப்பிகளாகும். எடுத்துக்காட்டாக நா இயக்க ஒலிப்பியாகும்.

இயங்கு ஒலிப்பி (active articulator)

ஒலிப்பாதையில் ஒலிகள் எழுப்பப்படும் போது நிலை ஒலிப்பிகளுடன் இயங்கி ஒலிகளை எழுப்பும் ஒலிப்பாதை உறுப்பு இயங்கு ஒலிப்பி எனப்படும். எடுத்துக்காட்டு நாக்கு; இது நிலை ஒலிப்பியமான அண்ணத்தைத் தொட்டு ன், ண், ட், ள் போன்றவை ஒலிகளை எழுப்புகிறது.

நிலை ஒலிப்பி (passive articulator)

ஒலிப்பாதையில் ஒரு ஒலி எழும் போது இயங்கா நிலையில் உள்ள ஒரு ஒலிப்பாதை உறுப்பு நிலை ஒலிப்பி எனப்படும். எடுத்துக்காட்டாக அண்ணம் நிலை ஒலிப்பி ஆகும். நாக்கு அண்ணத்தைத் தொட்டு ன், ண், ட் போன்ற ஒலிகளை எழுப்பும் போது நாக்கு மட்டுமே இயங்குகிறது. அண்ணம் இயங்குவதில்லை.

3.3.4. ஒலிப்பிடம் (Point of articulation)

ஒலிப் பாதையின் பல்வேறு பகுதியில் பல்வேறு உறுப்புகளால் ஒலியைத் தடுத்து நிறுத்தி பல்வேறு ஒலிகள் உண்டாக்கப்படுகின்றன. இந்த உறுப்புகள் ஒலிப்பிடம் எனப்படும். எடுத்துக்காட்டு இரு இதழ்கள் இணைந்து காற்றைத் தடுத்து நிறுத்துவதால் ப் என்ற ஒலி பிறக்கிறது. இது ஒரு ஒலிப்பிடம். ஒலி பிறக்கும் இடத்தை குறிப்பது ஒலிப்பிடம் எனப்படும். ஒலிப்பு என்பது மொழியில் சொற்கள் உச்சரிக்கப்படும் முறை அல்ல. மொழிக்கு உரிய ஒலிகளை வெளிப்படுத்தும் முறையாகும்.

இதழ் ஒலிகள் (Labial Sounds) - ப், ம் [p, m]

பல்லொலிகள் (Dental sounds) - த், ந் [t, n]

அண்பல்லொலிகள் (Alveolar sounds) – ற், ர், ன் [r, ɾ, n]

இடையண்ண ஒலிகள் (Palatal Sounds) – ச், ஞ், ய் [c, ɲ, y]

கடையண்ண ஒலிகள் (Velar Sounds) - க், ங் [k, ŋ]

நாமடி ஒலி (Refroflex Sounds) - ட், ண், ள், ழ் [ʈ, ɳ, ɭ, ɽ]

3.3.5. ஒலிப்பு முறை (manner of articulation)

ஒலிப்பு முறை அடிப்படையில் ஒலிகளை வெடிப்பொலி, உரசொலி, மருங்கொலி, ஆடொலி, மூக்கொலி, வருடொலி எனப் பகுக்கலாம். (பார்க்க மெய்யொலிகள்).

நுனி நா ஒலிப்பு (Apical Articulation)

நுனி நாவினால் நிலை உறுப்புகளைத் தொட்டு உண்டாக்கும் ஒலிப்பு நுனி நா ஒலிப்பு எனப்படும். எடுத்துக்காட்டாக ன், த் என்பதைக் கூறலாம்.

3.3.6. ஒலிப்பு வருணனை (Articulatory Description)

ஒலிகள் பிறக்கும் இடத்தையும் பிறக்கும் முறையையும் வர்ணிப்பது ஒலிப்பு வருணனை எனப்படும். மனிதனுடைய வாயறையில் உள்ள பல், நாக்கு, அண்ணம் போன்ற பல்வேறு இடங்களிலிருந்து பல்வேறு பேச்சொலிகள் பிறக்கின்றன. பேச்சொலிகள் பிறக்கும் வகையாலும்/முறையாலும் வேறுபடும். ஒரு ஒலி பிறக்கும் இடத்தையும் வகையையும் (முறையையும்) விவரித்தல் ஒலிப்பு வருணனை ஆகும். எடுத்துக்காட்டு ப் என்ற ஒலி இரு இதழ்கள் இணைந்து பிறக்கும் ஒலி;. இது பிறக்கும் முறையால் அடைப்பொலியாகும். இவ்வாறு ப் ஈரிதழ் அடைப்பொலி என வருணனை பெறும்.

உயிரொலி (vowel)

மூச்சுக்காற்றானது வாயறையில் எவ்விதத் தங்கு தடையுமின்றி அதாவது, இயங்கு ஒலிப்பான்களுக்கும் இயங்கா ஒலிப்பான்களுக்கும் இடையே தடையின்றிச் செல்லும்போது

உயிரொலிகள் பிறக்கின்றன. மெய்யொலிகள் பிறக்கும் போது முழுத்தடையோ சிறிதளவு தடையோ நிச்சயம் இருக்கும். உயிரொலிகள் மூன்று முதன்மையான பரிமாணங்கள் அடிப்படையில் வகைப்படுத்தப்படுகின்றன:

• வாய் திறந்திருக்கும் அளபு (மேல், மேலிடை, கீழிடை, கீழ் அல்லது திறந்த, பகுதிதிறந்த, பகுதி மூடிய, மூடிய)

• நாவின் உயர்ந்த பகுதியின் இருப்பிடம் (முன், நடு, பின்)

• இதழ்களின் இருப்பு (குவி, விரி)

மேற்கண்ட பரிமாணங்களின் அடிப்படையில் தமிழ் உயிர்களைப் பின்வருமாறு வகைப்படுத்தலாம்:

இ [i] - மேல் முன் இதழ் விரி குற்றுயிரொலி

ஈ [i:] - மேல் முன் இதழ் விரி நெட்டுயிரொலி

எ [e] - மேலிடை முன் இதழ் விரி குற்றுயிரொலி

ஏ [e:] - மேலிடை முன் இதழ் விரி நெட்டுயிரொலி

அ [a] - கீழ் நடு இதழ் விரி குற்றுயிரொலி

ஆ [a:] - கீழ் நடு இதழ் விரி நெட்டுயிரொலி

ஒ [டி] - மேலிடை பின் இதழ் குவி குற்றுயிரொலி

ஓ [o:] - மேலிடை பின் இதழ் குவி நெட்டுயிரொலி

உ [u] - மேல் பின் இதழ் குவி குற்றுயிரொலி

ஊ [u:] - மேல் பின் இதழ் குவி நெட்டுயிரொலி

தமிழில் பத்து உயிரொலிகளும் இரண்டு ஈருயிர்களும் உள்ளன.

பின் உயிரொலி (Back Vowel)

நாவின் பின்பகுதியை உயர்த்தி ஒலிக்கும் உயிரொலி பின் உயிரொலி எனப்படும். எடுத்துக்காட்டாக உ, ஒ. ஆகியவை பின்னுயிரொலிகள் ஆகும்.

முன் உயிரொலி (front Vowel)

நாவின் முன் பகுதியைத் உயர்த்தி ஒலிக்கும் உயிரொலி முன் உயிரொலி எனப்படும். நாவின் முன்பகுதி முன் எனப்படும். எடுத்துக்காட்டாக இ, எ ஆகியவை முன் உயிரொலிகள் ஆகும்.

ஈருயிர் (dipthongs)

ஓர் உயிர் ஒலியின் உச்சரிப்போடு தொடங்கி மற்றொரு உயிரொலியின் உச்சரிப்போடு முடிவடையும் உயிரொலியின் ஒலிப்பே ஈருயிர் (dipthongs) எனப்படுகிறது. இது ஒரு நெட்டுயிராக ஒலிக்கப்படுகிறது. இரண்டு உயிரொலிகளின் சாயலும் இந்நெட்டொலியில் காணப்படும். தமிழ் ஐ, ஔ ஆகிய இரண்டும் இவ்வாறு ஒலிக்கப்படும் ஈருயிர்களாகும். எ.கா. அ + இ = ஐ [ai] அ + உ = ஔ [au]

அரையுயிர் (semi vowels)

இது உயிரொலி போன்று உச்சரிக்கப்பட்டாலும் மெய்யொலி போன்று ஒலிப்பிடம் உடையதாக அமையும். எனவே மெய்யொலியின் தன்மையே அதிகம் பெற்றிருப்பதாகக் கருதப்படும். மேலும் உயிரொலி போன்று இது அசையாக (syllable) அமையாது. அதனால் ய், வ் இரண்டும் தமிழில் அரை உயிர்களாகக் கருதப்படுகின்றன. அவை முறையே இ, உ என்ற உயிர்களின் தன்மையைப் பெற்று விளங்குகின்றன.

குற்றுயிர்கள் (short vowel) நெட்டுயிர்கள் (long vowels)

ஒவ்வொரு ஒலியும் அவ்வொலிக்கே உரிய கால அளவோடு உச்சரிக்கப்படுகிறது. இக்கால அளவு மாத்திரை எனப்படும். சில ஒலிகளை உச்சரிக்கும் கால அளவு குறுகியும் சில ஒலிகளை உச்சரிக்கும் கால அளவு நீண்டு இருக்கும். ஒரு மாத்திரை கால அளவில் பிறக்கும் ஒலிகளை குறில் (short vowel) என்றும் இரண்டு மாத்திரை கால அளவில் பிறக்கும் ஒலிகளை நெடில் (long vowel) என்றும் வகைப்படுத்தலாம்.

விறைப்புயிர்கள் (tense vowels)

பேச்சுறுப்புகளின் தசையில் அதிகமான விறைப்பை ஏற்றி எழுப்பப்படும் உயிர்களுக்கு விறைப்புயிர்கள் என்று பெயர். உயிரொலிகளில் குறடில், நெடில் பாகுபாடில்லாத மொழிகளில் இவை விறைப்புயிர்களாக உச்சரிக்கப்படும் இயல்பைக் காணலாம். எடுத்துக்காட்டாக இந்தி மொழியில் [e, o, a] ஆகிய ஒலிகள் விறைப்புயிர்களாகும். தமிழ் மொழியில் நெட்டுயிர்களை விறைப்பாக உச்சரிக்கும் இயல்பு உண்டு. பிரெஞ்சு மொழியில் இவ்வகை உயிர்கள் மிகச் சாதாரணமாகக் காணப்படுகிறன.

நாவளை உயிர்கள் (Retroflex vowels)

உயிரொலிகளின் பிறப்பாக்கத்தில் இடைநா இடையண்ணத்தை நோக்கி எழும் போது முன்னுயிர்களும் இடைநாவிற்கும் கடைநாவிற்கும் இடையே உள்ள அண்ணப்பகுதியை நோக்கி எழும் போது நடுவுயிர்களும் பிறக்கின்றன. நா இவ்வாறு செயற்படும் போது நுனி நா மேற்பல்

வரிசையை நோக்கியோ அல்லது அண்பல்லை நோக்கி வளைந்தோ அல்லது இடையண்ணத்தை அடுத்து வளைந்தோ இருக்கும். இவ்வுயிர்கள் நாவளை உயிர்கள் எனப்படும்.

நெகிழ்வுயிர்கள் (Lax Vowels)

பேச்சுறுப்புகளின் தசைப்பகுதி தளர்ந்திருக்க எழுப்பப்படும் உயிர்களை நெகிழ்வுயிர்கள் எனலாம்.

குரலொலி (voiced sound)

குரலொலி என்பது குரல்வளை மடல்களனின் அதிர்வுடன் எழுப்பப்படும் ஒலி எனப் பொருள்படும். நுரையீரலிலிருந்து வெளியேறும் மூச்சு குரல்வளை மடல்களைக் கடந்த பின்னரே பல்வேறு ஒலிகள் பிறக்கின்றன. இக் குரல்வளை மடல்கள் ஓர் ஒலியை ஒலிக்க நெருக்கமாகக் கொணரும் போது அவை அதிர்கின்றன. இவ்வதிர்வோடு பிறக்கும் ஒலியைக் குரலொலி என்று கூறலாம்.

குரலிலா ஒலி (voiceless sound)

குரல்வளை மடல்களின் அதிர்வோடும் அதிர்வில்லாமலும் பிறக்கும் ஒலிகளை எளிதில் இனங்கண்டு கொள்ள முடியும். இரண்டு காதுகளையும் நன்றாகப் பொத்திக் கொண்டு ஆங்கில மொழியில் உள்ள *pack, bag* என்ற வார்த்தைகளை ஒன்றன் பின் ஒன்றாக உரக்க உச்சரித்தால் இரண்டாவது வார்த்தையில் தொடர்ச்சியாக ஒரு விதமான அதிர்வு இருப்பதை உணர முடியும். இவ்வதிர்வை விரல்களால் குரல்வளையை மென்மையாக அழுத்திப் பார்த்தும் உணரலாம். *pack* என்ற வார்த்தையை உச்சரிக்கும் போது இவ்வதிர்வு இருப்பதில்லை. மெய்யொலிகளில் அதிர்வொலிகளில் அதிர்விலா ஒலிகளுக்கு இணையாக அதிர்வொலிகள் உள்ளன.

மெய்யொலிகள் (consonants)

மூச்சுக்காற்று ஒலியுறுப்பால் தடுத்து நிறுத்தப்படுவதாலோ வேறுவகையில் மாற்றப்படுவதலோ மெய்யொலிகள் பிறக்கின்றன. அதாவது நுரையீரலிலிருந்து வெளியேறும் மூச்சு வாயறையில் அடைபடவோ அல்லது சிறிய இடுக்கின் வழியாகச் செல்லவோ அல்லது நாவின் இருமருங்கு வழியாகச் செல்லவோ அல்லது கடையண்ணம் மூக்கறை வாயிலைத் திறக்க மூக்கறையில் ஒலியூக்கம் பெற்று நாசி துவாரங்கள் வழியாக வெளியேறவோ செய்கையில் மெய்யொலிகள் பிறக்கின்றன. ஒலிப்பிடத்தையும் (இயங்கும் ஒலிப்பானும் இயங்கா ஒலிப்பானும் தொடர்பு கொள்ளுமிடம்) ஒலிப்பு முறையையும் அடிப்படையாகக் கொண்டு மெய்யொலிகளைத் தெளிவாக விளக்கலாம். இருப்பிடத்தை அடிப்படையாகக் கொண்டு மெய்யொலிகளைப் பின்வருமாறு பகுக்கலாம்.

இதழ் ஒலிகள் (Labial Sounds) - ப், ம் [p, m]

பல்லொலிகள் (Dental Sounds) - த், ந் [t, n]

அண்பல்லொலிகள் (Alveolar sounds) – ற், ர், ன் [r, r, n]

இடையண்ண ஒலிகள் (Palatal Sounds) – ச், ஞ், ய் [c, ɲ, y]

கடையண்ண ஒலிகள் (Velar Sounds) - க், ங் [k, ŋ]

நாமடி ஒலி (Refroflex Sounds) - ட், ண், ள், ழ் [t, ɳ, l, ɻ]

ஒலிப்பு முறை அடிப்படையில் ஒலிகளை வெடிப்பொலி, உரசொலி, மருங்கொலி, ஆடொலி, மூக்கொலி, வருடொலி எனப் பகுக்கலாம். இவ்வொலிகளின் வேறுபட்ட ஒலிப்பிடங்களையும் ஒலிப்பு முறைகளையும் அறிந்துகொள்வோம்.

அடைப்பொலி (stop or plosive)

இவற்றை வெடிப்பொலி (**plosive**) என்றும் கூறுவர். வாயறையின் ஒரிடத்தில் மூச்சுக்காற்று முழுவதும் தடை செய்யப்பட்டு திடீரென்று வெடிப்போடு வெளியேறினால் அதற்கு அடைப்பொலி என்று பெயர். இவ்வொலிப்பின் போது கடையண்ணம் மூக்கறையினுள் மூச்சு செல்லாதவாறு அடைக்கின்றது. குரல்வளை மடல்கள் அதிரவோ அதிராமலோ இருக்கலாம். தமிழில் வல்லின மெய்கள் அடைப்பொலிகளாகும். எ.கா. க், ச், ட், த், ப் [k, c, t, t, p].

மூக்கொலி (nasal)

அடைப்பொலி போலவே அமைந்து அண்ணக்கடை திறந்து மூக்கறை வழியே மூச்சுக்காற்று வெளியே செல்லும் போது மூக்கொலி தோன்றும். பெரும்பாலும் மூக்கொலி உண்டாகும்போது குரல்வளை மடல்கள் அதிர்கின்றன. தமிழில் மெல்லின மெய்கள் மூக்கொலிகளாகும்: ங், ஞ், ண், ந், ம் , ன் [ŋ, ɲ, ɳ , m, n].

உரசொலி (affricate)

வாயறையில் மூச்சுக்காற்று வரும்போது முழுவதும் தடுக்காமால் மூச்சு வெளியேறும் பாதையை ஒலிப்பானால் குறுக்கி அந்த இடுக்கின் வழியே செலுத்தினால் உராய்வுத்தன்மை (friction) ஏற்படும். இப்படி உராய்வுத் தன்மையோடு பிறக்கும் ஒலிகளே உரசொலிகள் எனப்படுகின்றன. தமிழில் சகரம் /c/ பாண்டி நாட்டுத் தமிழைத்தவிர பிற மாவட்டங்களில் /s/ என்று உரசொலியாகவே உச்சரிக்கப்படுகிறது. உரசொலி உண்டாகும் போது குரல்வளை மடல்கள் அதிர்ந்தும் அதிராமலும் இருக்கலாம். அப்போது அண்ணக்கடை அடைப்பும் ஏற்படும். பின்வருவன உரசொலிகளாகும்: [β, f, s, y, h].

மருங்கொலி (lateral)

மூச்சுக்காற்று வாயறையில் ஓர் இடத்தை அடைத்து நாக்கின் இரு மருங்கிலும் வெளியேறும் போது பிறக்கும் ஒலி மருங்கொலி எனப்படுகிறது. ல், ள், ழ் [l, ɭ, ʎ) மூன்றும் மருங்கொலிகளாகும். வருடொலி (flap) நாக்கின் நுனி மேலே எழுந்து உள்நோக்கி வளைந்து பின் வேகமாகக் கீழே வரும்போது அண்ணத்தில் மோதுவதால் எழுகிற ஒலியே வருடொலி. ஆடொலி (trill) மூச்சுக்காற்று வெளியேறும் போது இதழ்களோ நுனி நாக்கோ உள்நாக்கோ வேகமாக அதிர்ந்து ஆடும்படி அமைகிற ஒலியே ஆடொலி எனப்படுகின்றது. இன்று குமரி மாவட்ட மக்களால் மட்டுமே தமிழ் வழக்கில் ற் [r] என்ற ஆடொலி உச்சரிக்கப்பட்டு வருகிறது.

நுனி நா ஒலிப்பு (Apical Articulation)

நுனி நாவினால் அண்ணத்தைத் தொட்டு உண்டாக்கும் ஒலி நுனி நா அண்ணவொலி எனப்படும்.

3.3.7. தமிழ் ஒலிகளின் ஒலிப்பியல் ஆய்வு

தமிழ் ஒலிகளை ஒலிப்பு அடிப்படையில் உயிரொலிகள், மெய்யொலிகள் எனப்பகுத்து ஆயலாம்.

3.3,7.1. உயிரொலிகளின் ஒலிப்பியல் ஆய்வு

மூச்சுகாற்றானது வாயறையில் எவ்வித தங்கு தடையுமின்றி அதாவது, இயங்கு ஒலிப்பான்களுக்கும் இயங்கா ஒலிப்பான்களுக்கும் இடையே தடையின்றி செல்லும் போது உயிரொலிகள் பிறக்கின்றன. மெய்யொலிகள் பிறக்கும் போது முழுத்தடையோ சிறிதளவு தடையோ நிச்சயம் இருக்கும். உயிரொலிகள் மூன்று முதன்மையான பரிமாணங்கள் அடிப்படையில் வகைப்படுத்தப்படுகின்றன.

1.வாய் திறந்திருக்கும் அளபு (மேல், மேலிடை, கீழிடை, கீழ் அல்லது திறந்த, பகுதி திறந்த, பகுதி மூடிய, மூடிய)

2. நாவின் உயர்ந்த பகுதியின் இருப்பிடம் (முன், நடு, பின்)

3. இதழ்களின் இருப்புநிலை (குவி,விரி)

மேற்கண்ட பரிமாணங்களின் அடிப்படையில் தமிழ் உயிர்களை பின்வருமாறு பகுத்தாயலாம்.

இ – மேல் முன் இதழ் விரி குற்றுயிரொலி

ஈ – மேல் முன் இதழ் விரி நெட்டுயிரொலி

எ – மேலிடை முன் இதழ் விரி குற்றுயிரொலி

ஏ – மேலிடை முன் இதழ் விரி நெட்டுயிரொலி

அ – கீழ் நடு இதழ் விரி குற்றுயிரொலி

ஆ – கீழ் நடு இதழ் விரி நெட்டுயிரொலி

ஒ – மேலிடை பின் இதழ் குவி குற்றுயிரொலி

ஓ – மேலிடை பின் இதழ் குவி நெட்டுயிரொலி

உ – மேல் பின் இதழ் குவி குற்றுயிரொலி

ஊ – மேல் பின் இதழ் குவி நெட்டுயிரொலி

தமிழில் பத்து உயிரொலிகளும் இரண்டு ஈருயிரொலிகளும் உள்ளன.

ஈருயிர்

ஓர் உயிரொலியின் உச்சரிப்போடு தொடங்கி மற்றொரு உயிரொலியின் உச்சரிப்போடு முடிவடையும் உயிரொலியின் ஒலிப்பே ஈருயிர் எனப்படுகின்றது. இது ஓர் நெட்டுயிராக ஒலிக்கப்படுகின்றது. இரண்டு உயிரொலிகளின் சாயலும் இந்நெட்டொலியில் காணப்படும். தமிழ் ஐ, ஔ ஆகிய இரண்டும் இவ்வாறு ஒலிக்கப்படும் ஈருயிர்களாகும்.

அ + இ = ஐ

அ + உ = ஔ

3.3.7.2. மெய்யொலிகளின் ஒலிப்பியல் ஆய்வு

ஒலிப்பிடத்தையும் (இயங்கும் ஒலிப்பானும் இயங்கா ஒலிப்பானும் தொடர்புகொள்ளும் இடம்) ஒலிப்பு முறையையும் அடிப்படையாகக் கொண்டு தமிழ் மெய்யொலிகளை பகுத்தாயலாம். இருப்பிடத்தை அடிப்படையாகக் கொண்டு தமிழ் மெய்யொலிகளைப் பின்வருமாறு பகுக்கலாம்.

இதழ் ஒலிகள் – ப், ம்,

பல்லொலிகள் – த், ந்

அண்பல்லொலிகள் – ற், ர், ன்

இடையண்ண ஒலிகள் – ச், ஞ், ய்

கடையண்ண ஒலிகள் – க், ங்

நாமடி ஒலிகள் – ட், ண், ள், ழ்

ஒலிப்பு முறை அடிப்படையில் தமிழ் மெய்யொலிகளை வெடிப்பொலி, உரசொலி, ஆடொலி, மூக்கொலி, வருடொலி எனப் பகுத்தாயலாம்.

வெடிப்பொலி

வாயறையின் ஓரிடத்தில் மூச்சுக் காற்று முழுவது தடை செய்யப்பட்டு திடீரென்று வெடிப்போடு வெளியேறினால் அதற்கு வெடிப்பொலி என்று பெயர். அப்போது அண்ணக்கடை அடைப்பு ஏற்படும். தமிழில் வல்லின ஒலிகளே வெடிப்பொலிகளாகும்: க், ச், ட், ட், த், ப்.

மூக்கொலி

அடைப்பொலி போலவே அமைந்து அண்ணக்கடை திறந்து மூக்கறை வழியே மூச்சுக்காற்று வெளியே செல்லும் போது மூக்கொலி அல்லது மெல்லினம் தோன்றும். பெரும்பாலும் மூக்கொலி உண்டாகும் போது குரல் வளை மடல்கள் அதிர்கின்றன. பின்வரும் தமிழ் மெய்யொலிகள் மூக்கொலிகளாகும்: ங், ஞ், ண், ந், ம், ன்.

உரசொலி

வாயறையில் மூச்சுக்காற்று வரும்போது முழுவதும் தடுக்காமல் வெளியேறும் பாதையை ஒலிப்பானால் குறுக்கி அந்த இடுக்கின் வழியே செலுத்தினால் உராய்வுத் தன்மை ஏற்படும். இப்படி உராய்வுத்தன்மையுடன் பிறக்கும் ஒலிகளே உரசொலிகள் எனப்படுகின்றன. தமிழில் சகரம் (ச்) பாண்டிநாடுத் தமிழைத் தவிர் பிற மாவட்டங்களில் ஸ் என்று உரசொலியாகவே உச்சரிக்கப்படுகின்றது. உரசொலி உண்டாகும் போது குரல்வளை மடல்கள் அதிர்ந்தோ அதிராமலோ இருக்கலாகும்.

மருங்கொலி

மூச்சுக்காற்று வாயறையில் ஓர் இடத்தை அடைத்து நாக்கின் இரு மருங்கிலும் வெளியேறும் போது பிறக்கும் ஒலி மருங்கொலியாகும். தமிழில் ல், ள், ழ் என்பன மருங்கொலிகளாகும்.

வருடொலி

நாக்கின் நுனி மேலே எழுந்து உள்நோக்கி வளைந்து பின் வேகமாக கீழே வரும்போது அண்ணத்தில் மோதுவதால் எழுகிற ஒலியே வருடொலி எனப்படும். தமிழில் [ர்] வருடொலியின் கண் படும்.

ஆடொலி

மூச்சுக்காற்று வெளியேறும் போது இதழ்களோ நுனி நாக்கோ உள்நாக்கோ வேகமாக அதிர்ந்து ஆடும்படி அமைகிற ஒலியே ஆடொலி எனப்படுகின்றது. இன்று குமரிமாவட்ட மக்களால் மட்டுமே [ற்] என்ற ஆடொலி உச்சரிக்கப்படுகின்றது.

3.4. ஒலியியக்கவியல் ஆய்வு

ஒலியியக்கவியல் ஆய்வு என்பது பேச்சின் இயற்பியல் பண்புகளைப் (physical properties of speech) பற்றிய ஆய்வு ஆகும்; மேலும் பேச்சுக்குள் நிகழும் ஒலி அலை சமிக்ஞைகளை (sound wave signals) மாறுபட்ட நிகழ்வெண்கள் (frequencies), வீச்சுகள் (amplitudes) மற்றும் கால அளவுகள் (durations) மூலம் பகுப்பாய்வு செய்வதை நோக்கமாகக் கொண்டுள்ளது. இது முன்னர் கூறியது போல் ஒலியியல் என்பதன் துணைப் புலமாகும். இது பேச்சு ஒலிகளின் ஒலி

அம்சங்களைக் கையாளுகிறது. அலைவடிவத்தின் சராசரி சதுர அலைவீச்சு, அதன் காலம், அதன் அடிப்படை அதிர்வெண் போன்ற காலக் களப் பண்புக்கூறுகள் அல்லது அதிர்வெண் நிறமாலை (ஸ்பெக்ட்ரம்) போன்ற அதிர்வெண் களப் பண்புக்கூறுகள் அல்லது ஒருங்கிணைந்த நிறமாலை-காலம்சார் பண்புக்கூறுகள் மற்றும் ஒலிப்பியலின் பிற கிளைகளுடன் (ஒலிப்பியல் அல்லது கேட்பியல்) மற்றும் ஒலியன்கள், சொற்றொடர்கள் அல்லது கூற்றுகள் போன்ற அருவத்தன்மையான மொழியியல் கருத்துக்களுடன் இந்தப் பண்புகளின் உறவு இவற்றை ஒலிப்பியல் ஆராய்கிறது.

19ஆம் நூற்றாண்டின் பிற்பகுதியில் எடிசன் ஒலிப்பதிவுகருவி/ஃபோனோகிராஃப் (Edison phonograph) கண்டுபிடிப்பால் ஒலியியக்க ஒலியியல் ஆய்வு பெரிதும் மேம்படுத்தப்பட்டது. ஒலிப்பதிவுகருவி பேச்சு சமிக்ஞையை/குறிகையைப் பதிவு செய்ய அனுமதித்தது, பின்னர் செயலாக்க மற்றும் பகுப்பாய்வு செய்யப்பட்டது. ஒலிப்பதிவுக்கருவியிலிருந்து/ (ஃபோனோகிராஃபிலிருந்து) ஒரே பேச்சு சமிக்ஞையை/குறிகையை மீண்டும் மீண்டும் இயக்குவதன் மூலம், ஒவ்வொரு முறையும் வெவ்வேறு பேண்ட்-பாஸ் வடிப்பான் மூலம் வடிகட்டுவதன் மூலம், பேச்சு உரையின் ஸ்பெக்ட்ரோகிராம் கட்டமைக்கப்படலாம். 19 ஆம் நூற்றாண்டின் கடைசி இரண்டு தசாப்தங்களில் பிஃப்லெஜர்ஸ் காப்பகத்தில் வெளியிடப்பட்ட லுடிமர் ஹெர்மனின் தொடர் கட்டுரைகள் எடிசன் ஒலிப்பதிவுகருவியைப்/ஃபோனோகிராப்பைப் பயன்படுத்தி உயிரெழுத்துக்கள் மற்றும் மெய் எழுத்துக்களின் நிறமாலை பண்புகளை ஆராய்ந்தன, மேலும் இந்த ஆவணங்களில் தான் ஒலிச்செறிவு (formant/ஃபார்மண்ட்) என்ற சொல் முதலில் அறிமுகப்படுத்தப்பட்டது. வில்லிஸ் மற்றும் வீட்ஸ்டோனின் உயிர் உற்பத்தி கோட்பாடுகளுக்கு இடையில் வேறுபடுவதற்கு எடிசன் ஃபோனோகிராஃப் மூலம் செய்யப்பட்ட உயிரெழுத்து பதிவுகளையும் ஹெர்மன் மீண்டும் இயக்கினார்.

தொலைபேசித் துறையின் வளர்ச்சியால் ஒலியியக்கவியலின் அதிக முன்னேற்றம் சாத்தியமானது. (அலெக்சாண்டர் பெல்லின் தந்தை அலெக்சண்டர் மெல்விலெ பெல் ஒரு ஒலியல் ஆய்வாளர் ஆவார்). இரண்டாம் உலகப்போரில் (நிறமாலையைக் கண்டுபிடித்த) பெல் தொலைப்பேசிப் பரிசோதனைக் கூடங்கள் (Bell Telephone Laborataries) காலநீட்சி மாறாத (periodic) மற்றும் காலநீட்சி மாறும் (aperiodic) பேச்சு ஒலிகளின் நிறமாலைசார் பண்புகள், பேச்சுக்குழல் ஒலியூக்கங்கள் மற்றும் உயிர் ஒலிச்செறிவுகள், குரல் தரம், மீக்கூறு போன்றவற்றின் முறையான ஆய்வுக்கு வசதிசெய்தது.

ஒருங்கிணைந்த நேரியல் முன்கணிப்பு எச்சங்கள் (Integrated linear prediction residuals (ILPR/ஐ.எல்.பி.ஆர்) 1995ஆம் ஆண்டில் டி வி அனந்தபத்மநாபாவால் (Ananthapadmanabha 1995) முன்மொழியப்பட்ட ஒரு சிறந்த அம்சமாகும். இது குரல் மூல சமிக்ஞையை/குறிகையை நெருக்கமாக மதிப்பிடுகிறது. இது சகாப்தங்களின் அல்லது தொண்டை மூடல் நேர்வின் துல்லியமான மதிப்பீட்டில் (Prathosh et al 2013) மிகவும் பயனுள்ளதாக இருந்தது. எ.ஜி ராமகிருஷ்ணன் மற்றும் பலர் (Ramakrihshna et al 2015) ஐ.எல்.பி.ஆரின் தனித்துவமான கொசைன் உருமாற்றக் குணகங்களில் (cosine transform coefficients) மெல் அதிர்வெண் செப்ஸ்ட்ரல் குணகங்களுக்கு (ceptral coefficients) துணைபுரியும் பேசுபவர் தகவல்கள் உள்ளன என்பதை 2015இல் காட்டினர். நிறுத்த மெய்யின் மூடல்-வெடிப்பு மாற்றத்தைப் பண்பாக்கம் செய்வதற்காக டி வி அனந்தபத்மநாபா மற்றும் பலர் (Ananthapadmanabha 2015) அறிமுகப்படுத்திய மற்றொரு அளவீட்டு, நேர-கள அம்சமாகும்.

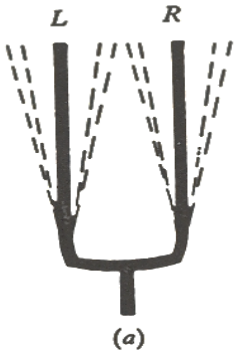
கோட்பாடு மட்டத்தில், பேச்சு ஒலியியக்கவியல் மின் சுற்றுகளுக்கு ஒத்ததாக வடிவமைக்கப்படலாம். புதிய மின்சாரக் கோட்பாட்டை (electric theory) ஒலியியக்கவியலில் பயன்படுத்த முடியும் என்பதை முதன்முதலில் அங்கீகரித்தவர்களில் லார்ட் ரேலேயும் (Lord Rayleigh) இருந்தார்; ஆனால் 1941ஆம் ஆண்டு வரை சுற்று மாதிரி (circuit model) திறம்பட பயன்படுத்தப்படவில்லை என்று சிபா மற்றும் கஜியாமா (Chiba and Kajiyama) எழுதிய "உயிரொலி: அதன் இயல்பு மற்றும் அமைப்பு" (The Vowel: Its Nature and Structure) நூலிலிருந்து தெரிகிறது. (இந்த நூல் ஆங்கிலத்தில் ஜப்பானில் பணியாற்றிய ஜப்பான் நூலாசிரியர்களால் உலகப்போரின் உச்சத்தில் வெளியிடப்பட்டது,) 1952இல் ரோமன் ஜாகப்சன், குன்னார் ஃபாண்ட் மற்றும் மோரிஸ் ஹல்லே ஆகியோர் “பேச்சு ஆய்வின் முதனிலை ஏற்பாடுகள்) (Preliminaries to Speech Analysis) ஒன்றாக கோட்பாடு என்ற நூலை எழுதினர்; இது ஒலியியக்கவியலையும் ஒலியனியல் கோட்பாட்டையும் ஒன்றுசேர்க்க முயற்சித்தது. இந்தச் சிறிய புத்தகத்தைத் தொடர்ந்து 1960ஆம் ஆண்டில் ஃபான்ட் (Fant) "பேச்சு உற்பத்தியின் ஒலியியக்க கோட்பாடு" (Acoustic Theory of Speech Production) என்ற நூலை எழுதினார்; இது கல்வித்துறை மற்றும் தொழில் துறை இரண்டிலும் பேச்சு ஒலியியக்க ஆராய்ச்சிக்கான முக்கிய கோட்பாட்டு அடித்தளமாக இருந்து வருகிறது. (தொலைபேசித் துறையில் ஃபான்ட் மிகவும் ஈடுபாடு கொண்டிருந்தார்.) இந்தத் துறையில் பிற முக்கியமான பங்களிப்பாளர்கள் "ஒலியியக்கவியல்", (Acoustic Phonetics) எழுதிய கென்னத் என். ஸ்டீவன்ஸ் (Kenneth N. Stevens, 2000) மற்றும்

ஒசாமு புஜிமுரா மற்றும் பீட்டர் லேட்ஃபோகெட் (Osamu Fujimura, and Ladefoged) ஆகியோர் ஆவர்.

ஒலியியக்கவியல் என்பது பேச்சின் இயற்பியல் பண்புகளைப் பற்றிய ஆய்வு ஆகும், மேலும் பேச்சுக்குள் நிகழும் ஒலி அலை சமிக்ஞைகளை மாறுபட்ட அதிர்வெண்கள், வீச்சுக்கள் மற்றும் கால அளவுகள் மூலம் பகுப்பாய்வு செய்வதை நோக்கமாகக் கொண்டுள்ளது. பேச்சு ஒலிகளின் ஒலியியல் பண்புகளை நாம் பகுப்பாய்வு செய்ய ஒரு வழி அலைவடிவத்தைப் பார்ப்பது. அழுத்த மாற்றங்களை (pressure changes) ஒரு அலைவடிவத்தில் வரையாலாம்; இது காற்று துகள்கள் சுருக்கப்படுவதையும் மற்றும் அடர்த்தி குறைவதையும் எடுத்துக்காட்டுகிறது; இது வெளிப்புறமாக பரவும் ஒலி அலைகளை உருவாக்குகிறது. ஒரு இசைக்கவடி (ட்யூனிங் ஃபோர்க், tuning fork) தாக்கப்படுவது காற்றில் உள்ள அழுத்தம் ஏற்ற இறக்கங்களுக்கும், ஒலியை நாம் உணரும்போது காற்று துகள்கள் எவ்வாறு ஊசலாடுகின்றன (ஒரு திசையில் தாளமாக நகரும்) என்பதற்கும் ஒரு எடுத்துக்காட்டு ஆகும்.

அலைவடிவு (waveform)

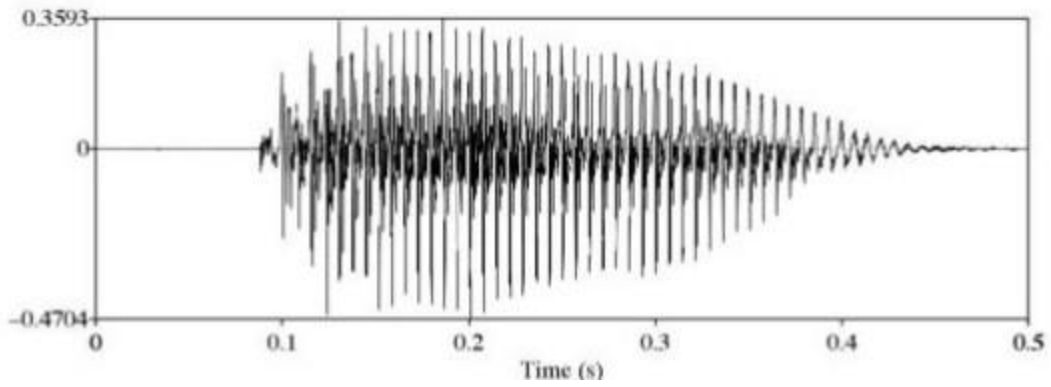
பேச்சு ஒலிகளின் ஒலியியக்கப் பண்புகளை நாம் பகுப்பாய்வு செய்ய ஒரு வழி அலைவடிவத்தைப் பார்ப்பது ஆகும். அழுத்த மாற்றங்களை ஒரு அலைவடிவத்தில் வரையலாம்; இது வெளிப்புறமாகப் பரவும் ஒலி அலைகளை உருவாக்கும் காற்று துகள்கள் சுருக்கப்படுவதையும் அடர்த்திகுறைவதையும் எடுத்துக்காட்டுகிறது. இது ஒரு இசைக்கவடி தாக்கப்படும்போது காற்றில் உள்ள அழுத்தத்தின் ஏற்ற இறக்கங்களையும் ஒலியை நாம் உணரும்போது (ஒரு திசையில் தாளமாக நகரும்) காற்று துகள்கள் எவ்வாறு ஊசலாடுகின்றன என்பதற்கும் எடுத்துக்காட்டைத் தரும்.



(a)

(b)

உயிரொலியின் அலைவடிவம்: Ogden 2009: 30



அதிர்வெண், அலைவீச்சு (Frequency vs Amplitude)

ஒரு ஒலியின் அதிர்வெண் (இசைமை/pitch) மற்றும் வீச்சு ('சத்தம்'/loudness அல்லது தீவிரம்/intensity) ஒரு அலைவடிவத்தில் பகுப்பாய்வு செய்யலாம். மீண்டும் மீண்டும் நிகழும் அமைப்பொழுங்குடன் காலம் மாறா அலைவடிவத்தில் சுழற்சிகளின் எண்ணிக்கையின் மூலம் அதிர்வெண்ணைக் கணக்கிடலாம். வினாடிக்கு அதிக சுழற்சிகளின் எண்ணிக்கையானது அதிக நிகழ்வெண்ணையும் உணரப்படும் இசைமையையும் கொண்டிருக்கும். நிகழ்வெண் பொதுவாக ஹெர்ட்ஸ்-இல் (Hertz- Hz) வெளிப்படுத்தப்படுகிறது.

ஒரு அலைவடிவத்தின் வீச்சுகளைப் பகுப்பாய்வு செய்வது ஒரு ஒலி எவ்வளவு தீவிரமானது/கடுமையானது அல்லது 'சத்தமாக' இருக்கிறது என்பதையு காற்று துகள்கள் எவ்வளவு விலகும் என்பதையும் கூறுகிறது. இது வழக்கமாக டெசிபல்களில் (dB/டி.பி.) வெளிப்படுத்தப்படுகிறது. அலைவடிவத்தில் உள்ள எக்ஸ்-அச்சு ஒலி உருவாக்கப்பட்ட கால எல்லைக்கு ஒத்திருக்கிறது (வழக்கமாக விநாடிகள் அல்லது மில்லி விநாடிகளில்), மற்றும் y- அச்சு வீச்சுகளைக் குறிக்கிறது.

சைன் அலைகள், கலவை அலைகள் (sine waves and complex waves)

சைன் அலைகள் (sine waves) மிகவும் எளிமையான, வழக்கமாக மீண்டும் மீண்டும் வரும் வடிவங்களைக் கொண்ட அலைவடிவங்கள் ஆகும். அலைவடிவத்தில் உள்ள 'சுழற்சிகளின்' எண்ணிக்கை (கால அலைவடிவத்தின் முழுமையான மீள்நிகழ்வு எண்ணிக்கை) காட்டப்படும் கால எல்லைக்குள் குரல்வள மடல்கள் எத்தனை முறை திறக்கப்பட்டுள்ளன என்பதைப் பிரதிபலிக்கிறது. இது அடிப்படை அதிர்வெண் (f0) என அழைக்கப்படுகிறது; இது ஹெர்ட்ஸ் (Hz)இல் அளவிடப்படுகிறது. 200 ஹெர்ட்ஸ் அதிர்வெண் என்றால் அலைவடிவத்திற்குள் வினாடிக்கு 200 நூறு முழுமையான சுழற்சிகள் உள்ளன; எனவே குரல்வளை மடல்கள் 200

தடவைகள் திறக்கப்பட்டுள்ளன. உண்மையில், பெரும்பாலான பேச்சு ஒலி அலைகள் மிகவும் கலவைத்தன்மையான வடிவத்தைக் கொண்டுள்ளன, மேலும் அவை கலவைத்தன்மையான அலைகள் (complex waves) என்று அழைக்கப்படுகின்றன. இவை இரண்டு அல்லது அதற்கு மேற்பட்ட எளிய சைன் அலைகளால் ஆனவை; மேலும் ஒரு அலைவடிவத்தில் வினாடிக்குச் சுழற்சிகளின் எண்ணிக்கையைக் கணக்கிடுவதன் மூலம் அடிப்படை நிகழ்வெண்ணை கலவைத்தன்மையான அலைவடிவங்களிலும் கணக்கிடப்படலாம்.

கால நீட்சி மாறா ஒலி அலைகள், கால்நீட்சி மாறும் ஒலி அலைகள் (Periodic vs Aperiodic sound wave)

சைன் மற்றும் கலவைத்தன்மையான அலைவடிவங்கள் காலநீட்சிமாறா அலைகளாகும் (periodic waves); அதாவது அவற்றின் சுழற்சிகள் ஒழுங்கானவை மற்றும் மீண்டும் மீண்டும் நிகழ்கின்றவை. காலநீட்சிமாறா அலையாகத் தோன்றும் பேச்சு ஒலிகளின் வகைகள் உயிரெழுத்துக்கள் அல்லது மூக்கொலிகள் போன்ற குரல் ஒலிகள் ஆகும். இத்தகைய ஒலிகள் ஒழுங்காக மீண்டும் மீண்டும் நேரும் அலைவடிவங்களைக் கொண்டிருப்பதால், அவை சைன் அலைகளை உடைக்கும் 'ஃபோரியர் பகுப்பாய்வு' ('Fourier analysis) மூலமாகவும் குறியத்திறவு செய்யலாம். இந்த வகை வரைபடம் நிறமாலை/ஸ்பெக்ட்ரம் என்று அழைக்கப்படுகிறது, இது நேரத்தை அளவிடாது. அதற்கு பதிலாக, x-அச்சு நிகழ்வெண்ணை அளவிடும், மற்றும் y-அச்சு ஒலி அழுத்த அளவைக் உருப்படுத்தம் செய்யும்.

இந்த வகை அலைகளின் குறைந்த அதிர்வெண்/நிகழ்வெண் கூறுகளைத் தேர்ந்தெடுப்பதன் மூலம் இந்த வகை வரைபடத்தின் அடிப்படை அதிர்வெண்/நிகழ்வெண் உருவாக்கப்படலாம். இது வழக்கமாக நிறமாலையின்/ஸ்பெக்ட்ரமின் முதல் முழுமையான உச்சமாகும். இந்த அடிப்படை அதிர்வெண் உச்சத்திலிருந்து, ஒலியையபுகள்/ஹார்மோனிக்ஸ் (harmonics) சம இடைவெளியில் உள்ள முழு எண் பெருக்கங்களில் நிகழ்கிறது. ஒலியியையபுகள்/ஹார்மோனிக்ஸ் குரல்வளையில் உள்ள 'இயற்கை ஒலியூக்கம்' (natural resonance) என அழைக்கப்படுகிறது; அவை பெருக்கப்பட்ட அதிர்வெண்கள் ஆகும். நிறமாலையில்/ஸ்பெக்ட்ரமில், இவை ஒவ்வொரு உச்சத்தையும் ஒத்திருக்கும்.

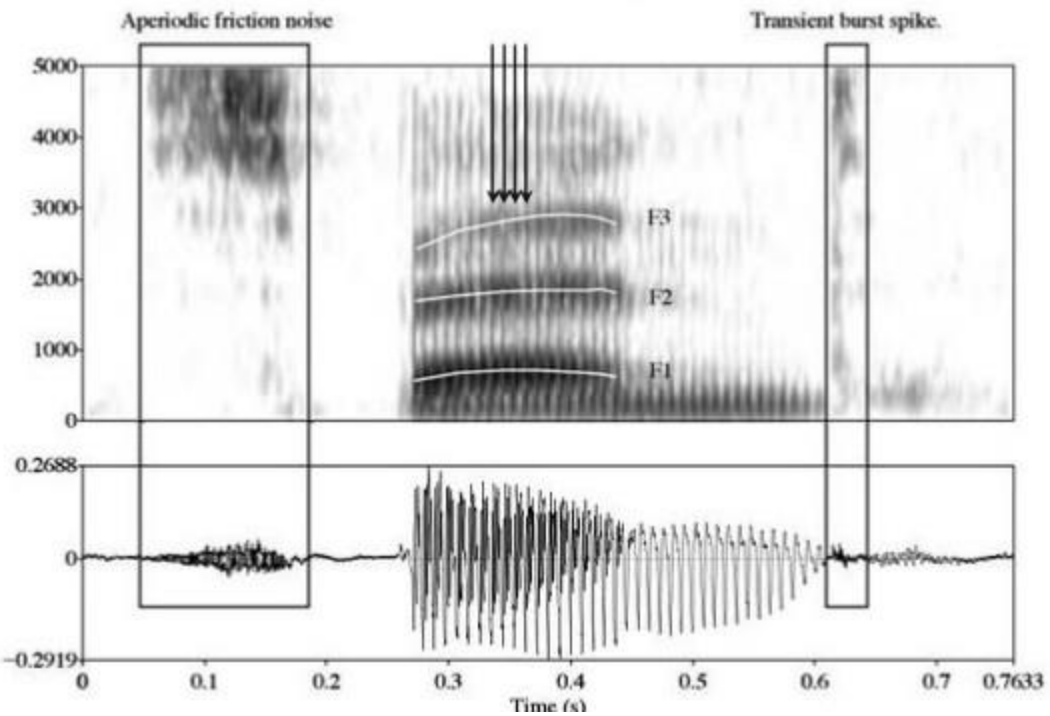
மாறாக, பேச்சு ஒலிகளை ஒலியியல் ரீதியாக பகுப்பாய்வு செய்யும் போது அவை காலநீட்சி மாறுபவை ஆகவும் இருக்கும். இதன் பொருள் என்னவென்றால், அவை ஒழுங்கான தொடர்ச்சியான முறையைக் கொண்டிருக்கவில்லை; மாறாக, அவை மிகவும் சீரற்ற வடிவத்தைக் கொண்டிருக்கின்றன; அதாவது ஒரு அடிப்படை அதிர்வெண்ணைக் கணக்கிட முடியாது. இதன்

பொருள், காலநீட்சி மாறும் பேச்சு ஒலிகள் குரலற்றவை; அதாவது குரலற்ற உரசொலி போன்றவை.

நிறமாலைகள் (Spectrograms)

ஒரு ஒலியை ஒலியியக்கவியல் ரீதியாக பகுப்பாய்வு செய்வதற்கான மற்றொரு வழி நிறமாலை/ஸ்பெக்ட்ரோகிராம் பார்ப்பதன் மூலம் ஆகும். அலைவடிவத்தில் நாம் காணக்கூடியதை விட அவை மிகவும் சிக்கலான தகவல்களை வழங்குகின்றன. அலைவடிவங்களைப் போலவே, நேரம் x- அச்சில் காட்டப்படும், ஆனால் y-அச்சு ஒலியின் அதிர்வெண்ணை அளவிடுகிறது. ஒலியியக்க ஆற்றலில் கருமையால்/இருளினால் வீச்சு குறிப்பிடப்படுகிறது. சத்தமாக ஒலி இருந்தால் நிறமாலையில்/ஸ்பெக்ட்ரோகிராமில் அதிக கருமையுடன் தோன்றுகிறது, எனவே இது மிகவும் தீவிரமானது/கடுமையானது. நிறமாலைகள்/ஸ்பெக்ட்ரோகிராம்கள் காலநீட்சி மாறும் ஒலிகளுடன் வரும் அதிக அதிர்வெண் ஆற்றலைக் காண அனுமதிக்கின்றன.

படம்: ஒரு சொல்லின் நிறமாலை; காலநீட்சி மாறா, காலநீட்சி மாறும், கடத்தல் ஓசைகள் அடையாளப்படுத்தப்பட்டுள்ளன (Ogden 2009: 32)



நிலைமாற்ற ஒலிகள், தொடர்ச்சியான ஒலிகள் (Transients vs Continuous sounds)

நிலைமாற்றங்கள் என்பது காலநீட்சி மாறும் ஒலியின் ஒரு வடிவம். இந்த ஒலி ஒரு மூடலின் பின்னால் அழுத்தத்தை உருவாக்கும் பின்னர் அலைவடிவத்தில் ஒரு கூர்மையான முனையாக தோன்றுகின்ற திடீர் வெடிப்பை/வெளியீடைக் கொண்டிருக்கும். ஒரு நிறமாலையில்,

வெளியீட்டைக் குறிக்க ஒலியியக்க ஆற்றலின் கருமையான செங்குத்து பட்டைக்கு முன் மூடல் ஒரு வெற்று இடமாகக் காட்டப்படுகிறது. ஒரு வழக்கமான நிலைமாற்ற ஒலி ஆங்கில ஒலி ஒழுங்குமுறையில் [p] அல்லது [b] போன்ற ஒரு வெடிப்பு ஒலிகளாகும்.

தொடர்ச்சியான ஒலிகள் காலநீட்சி மாறும் ஒலியின் மற்றொரு வடிவம். குரல் உயிர் ஒலியைப் போலன்றி, இது அலைவடிவத்தில் ஒழுங்கற்ற, சீரற்ற வடிவமாகக் காண்பிக்கப்படும். நிறமாலையில், இது இது கருமையான மற்றும் கடுமையான உயர் அதிர்வெண் ஒலி ஆற்றலால் குறிக்கப்படும்; எனவே அதிக வீச்சு உள்ளது. இது ஆங்கில ஒலி ஒழுங்குமுறையில் [f] மற்றும் [s] போன்ற குரலற்ற உரசொலிகள் ஆகும்.

நிறமாலையில் குரல்

முன்னர் நிறுவப்பட்டபடி, குரல் ஒலி காலநீட்சிமாறா குறிகைகள்/சமிக்ஞைகள் ஆகும். குரல்வளை மடல்கள் அதிர்வுறும் போது ஒழுங்கான திறப்புகளின் காரணமாக ஒரு அடிப்படை அதிர்வெண்ணைக் கணக்கிட முடியும். ஒரு அலைவடிவத்தில், இது ஒரு காலநீட்சிமாறா அலை மூலம் முன்னிலைப்படுத்தப்படும். ஒரு நிறமாலையில், கவனிக்க இரண்டு குறிப்பிட்ட காட்சி கூறுகள் (visual elements) உள்ளன; அவை குரல் ஒலியை ஒத்திருக்கின்றன. முதலாவது செங்குத்தான வரிகள் (vertical striations) (அவை நிறமாலையில் செங்குத்து அலை அலையான கோடுகள் போல இருக்கும்); இது ஒவ்வொரு முறையும் காற்று குரல்வளை மடல்களின் வழியாகப் பாயும் போது குரல்வளை மடல்களின் திறப்புடன் பொருந்தும்; மேலும் மற்ற காட்சி துப்பு/குறிப்பு (visual clue) உயிரெழுத்துக்கள், அரை உயிர்கள் மற்றும் மூக்கொலிகளின் பொதுவான கருமையான கிடைமட்ட பட்டைகள் ஆகும். இவை ஒலிச்செறிவுகள் (formants) என்று அழைக்கப்படுகின்றன; அவை பேச்சுக்குழலின் இயல்பான அதிர்வுகளாகும் (முன்பு, அவை ஒலியைபுகள் என விவரிக்கப்பட்டன). ஒலிச்செறிவு மாறுவதை அனுமதிக்கும்படி பேச்சுக் குழலின் பருமனையும் வடிவத்தையும் மாற்றியமைக்கலாம். நாவின் நிலை, உதட்டின் நிலை போன்றவற்றை மாற்றுவதன் மூலமும் இதைச் செய்யலாம்.

3.4.1. தமிழ் ஒலியியக்கவியல் ஆய்வுகள்

இப்பகுதியில் தமிழ் ஒலிகளின் ஒலியியக்கவியல் ஆய்வு ஒலிவர்ணனைப் படத்தின் உதவியுடன் விளக்கப்படும். தமிழ் ஒலியியக்கவியல் குறித்த முருகையன் மற்றும் இரவிசங்கரின் ஆய்வுகள் மிக முக்கியமானவையும் அடிப்படையானவையும் ஆகும். இரவிசங்கரின் தமிழ் இசையோட்டம் குறித்த ஆய்வு (Ravisankar 1980, 1988, 1994) உரையிலிருந்து பேச்சாகத்திற்கு இயற்கைத் தன்மையை நல்கப் பனுள்ளதாய் அமையும். கருவி ஒலியியக்கவியல்

அடிப்படையிலான யக்யநாராயணா, ஹேமா மூர்த்தி, ராஜேந்திரன் ஆகியோரின் ஆய்வுகளும் நல்ல நோக்கீடுகளாக அமையும்.

தமிழ் ஒலிகளின் ஒலியியக்கவியல் ஆய்வுகள் மேற்கொள்ளப்பட்டு கட்டுரைகள் பல வெளிவந்துள்ளன (பார்க்க நோக்கீட்டுநூல்கள்). சில ஆய்வுகள் குறிப்பிட்ட தமிழ் ஒலிகளின் ஒலியியக்கவியல் குறித்து ஆய்ந்தன. எடுத்துக்காட்டாக [ழ்], [ள்] என்ற ஒலிகளின் ஒலியியக்கவியல் ஆய்வுகள் மேற்கொள்ளப்பட்டு பல உண்மைகள் வெளிவந்துள்ளன (Narayan et al 1999). நடனசபாபதியின் தமிழ் ஒலியிக்கவியல் ஆய்வுகளும் (Natanasapapathy 1974, 2002) நல்ல நோக்கீடுகளாக அமையும். கார்த்திகா என்பவரால் தமிழ் ஒலிகளின் ஒலியியக்கவியல் ஆய்வு ஒரு ஆய்வுத்திட்டமாக மேற்கொள்ளப்பட்டு வருவதாக இணையதளம் மூலம் அறியமுடிகின்றது. ஸ்ரீநாத் மற்றும் பிறர் (Srinatha et al 1999) தமிழ் ல், ள், ர், ற், ழ் ஆகிய ஒழுகொலிகள் பற்றி ஒலியியக்கவியல் அடிப்படையில் ஆய்வு செய்துள்ளனர். அவர்கள் ஒழுகொலிகளின் ஒலிப்பு-ஒலியிக்க உறவுகள் பற்றி ஆய்ந்துள்ளனர். யக்யநாராயணா, ஹேமா மூர்த்தி ஆகியோரின் ஆய்வுகளும் (Hema Murthy and Yegnanarayana 1991) இங்கு குறிப்பிடத் தகுந்தன.

3.4.2. தமிழ் ஒலியியக்கவியல் பற்றிய ஒரு எடுத்துக்காட்டான ஆய்வு

தமிழ் ஒலியன்களின் ஒலியியக்கப் பண்புக்கூறுகள் இவ்வாய்வில் ஆயப்படுகின்றன (EEN540-Project1- The Acoustic Features of Speech Sounds & Frequency Domain Filtering). தமிழில் 45 ஒலியன்கள் உள்ளன; இவை உயிர்களையும், மெய்களையும் உயிர் ஒன்றிணைப்புகளையும் உட்படுத்தும். இவ்வொலியன்களைக் கொண்டிருக்கும் எல்லா சொற்களும் 3D Sound Blaster Pro in G-60 Hewlett Packard-இல் இணைக்கப்பட்டுள்ள லாப்டெக் ஒலிபெருக்கியைப் (Labtech Microphone) பயன்படுத்தி ஒலிப்பதிவு செய்யப்பட்டுள்ளன. கேட்பை பிடிக்க துடுக்கு பயன்படுத்தப்பட்டது மற்றும் எல்லா சொற்களும் ஓரொலியாக 16KHz மாதிரி விகிதத்துடன் மற்றும் 16 பைட் ஆழத்துடன் பதிவு செய்யப்பட்டன.

.wav கோப்பை மட்லாபுக்குள் (MATLAB) படித்தவுடன் பிளாட் கட்டளை ஆயப்பட்டுள்ள சொல்லின் காலப் பொருண்மைக்கள அலைவடிவைப் படம் வரைய பயன்படுத்தப்படுகின்றது.

ஒலிநிறமாலை வரைவு/ஒலியியக்க வர்ணனைப்படம் (spectrogram) எவ்வாறு கால அடிப்படையில் குறிகையின் வண்ணப்பட்டை அடர்த்தி வேறுபடுகின்றது என்பதைக் காட்டும் தோற்றமாகும். சொற்களின் குறுகிய மற்றும் அகன்ற வண்ணப்பட்டைப் பதிவு நிழற்படங்கள் ஆயப்பட்டு காட்டப்படுகின்றது. குறுகிய பட்டை நிறமாலை வரைவுகள் (narrow band spectrograms) கிடைமட்டமான ஒத்திசைகளும் அகன்ற பட்டை நிறமாலை (wide band

spectrograms) வரைவுகள் செங்குத்து ஒத்திசைகளும் (vertical harmonics) உள்ளன என்பது கவனத்தில் கொள்ளப்படுகின்றது. ஒலிநிறமாலை வரைவின்/ஒலியியக்க வர்ணனைப்படம் முப்பரிமாண நிறமாலை வரைபடமும் (three dimensional plot of the spectrogram) காட்டப்படுகின்றது.

ஒலியனைக் கொண்டிருக்கும் சொல்லின் பகுதி கருத்தில் கொள்ளப்படுகின்றது மற்றும் இப்பகுதியின் பரும அளவு நிறமாலை (magnitude spectrum) டெசிபெல்லில் வரையப்படுகின்றது.

பரும நிறமாலை வரைவு/ஒலியியக்க வர்ணனைப்படம் நேர்கோடான ஊக நிறமாலைசார் உறையுடன் மேல் சுமத்தப்பட்டுள்ளது. ஒவ்வொரு சொல்லின் பகுதியின் நீளம் வேறாகும். எனவே சொல்லில் ஒலியன் வரும் இடங்கள் குறிக்கப்பட்டுள்ளன மற்றும் பரும நிறமாலை நேர்கோடான ஊக நிறமாலைசார் உறை (Linear Prediction Spectral Envelope) வரையப்படுகின்றது.

பகுப்பாய்விற்குப் பயன்படுத்தப்படும் சாளரம் ஒரு ஹமினிங் சாளரம் (hamming window) மற்றும் ஃபாஸ்ட் ஃபோரியர் உருமாற்றத்தைக் (Fast Fourier Transform) கணக்கிடுவதற்கான பின்களின் (bins) எண்ணிக்கை 1024 ஆகும்.

இவ்வாய்வில் பயன்படுத்தப்பட்டுள்ள சொற்கள் கீழே பட்டியலிடப்பட்டுள்ளன. ஒவ்வொரு சொல்லும் கால அலைவடிவம், குறுகிய பட்டை நிறமாலை வரைவு, அகன்ற பட்டை நிறமாலை வரைவு/ஒலியியக்க வர்ணனைப்படம், நிறமாலையின்/ஒலியியக்க வர்ணனைப்படம் முப்பரிமாண வரைவு, ஆய்வுக்குள்ளாகும் ஒலியனின் பரும வரைவு மற்றும் ஒன்றின் மேல் ஒன்று வரும் பருமம் மற்றும் நேர்கோடான ஊக நிறமாலை உறையைக் காட்டும் வரைவு. சீரான வளைவு வரைபடத்தில் கறுப்பு நிறத்தில் காட்டப்பட்டுள்ளது. மேலும் சொல் ஒவ்வொரு பக்கத்திலும் ஒன்றிணைக்கப்பட்டுள்ளவைகளைச் சுட்டுவதால் சொல்லைக் கேட்க இயலும்.

தமிழ் ஒலிகளை பிராத் போன்ற மென்பொருளைப் பயன்படுத்தி ஒலியியக்கப் பின்னணியில் ஆய இயலும். பிராத் தமிழ் ஒலிகளின் ஒலியியக்கக் பண்புக்கூறுகளை நிறமாலை நிழற்படங்களாகத் தரவல்லது. இந்நிறமாலை நிழற்படங்கள் மூலம் தமிழ் உயிரொலிகள் மற்றும் மெய்யொலிகள் ஒலியியக்க அடிப்படையில் ஆயப்பட்டு ஆய்வு முடிவுகள் தரப்பட்டுள்ளன.

ஒலிப்பு முறையில் ஏற்படும் வேறுபாடுகளை ஒலியியக்கப் பண்புக்கூறுகளின் வேறுபாடுகளுடன் தொடர்புபடுத்திப் பார்க்க இயலுகின்றது. தமிழ் உயிரெழுத்துக்கள் அவற்றில் ஒலிச்செறிவு அமைப்பால் வேறுபடுவதை அறிய முடிகின்றது. உயிரெழுத்துகளின் ஒலிச்செறிவு அமைப்பு அவை வரும் சூழல் அடிப்படையில் மாறுவதையும் அறிய முடிகின்றது.

மெய்யெழுத்துக்கள் அவற்றின் ஒலிப்பிடம் மற்றும் ஒலிப்பு முறை அடிப்படையில் வேறுபடுவதன் காரணமாக அவ்வேறுபாடு அவற்றின் ஒலியியக்கப்பண்புக்கூறுகள் அடிப்படையிலும் வேறுபடுகின்றன. ஒலிப்பிடங்களையும் அவற்றுடன் தொடர்புள்ள ஒலியியக்கப் பண்புக்கூறுகளையும் ஒலிப்புமுறைகளையும் அவற்றுடன் தொடர்புள்ள ஒலியியகக்கப் பண்புக்கூறுகளையும் தொடர்புபடுத்திப் பார்க்க இயலுகின்றது.

3.4.2.1 தமிழ் உயிர்களின் ஒலியியக்கவியல் ஆய்வு

தமிழில் *இ, ஈ, உ, ஊ, எ, ஏ, ஒ, ஓ, அ, ஆ, ஐ, ஔ* என்ற உயிரொலிகள் உள்ளன. இதில் *இ* என்பது மேல் முன் இதழ் விரி குற்றுயிரொலியாகும்; *ஈ* என்பது மேல் முன் இதழ் விரி நெட்டுயிராகும். உ என்பது மேல் பின் இதழ்குவி குற்றுயிரொலியாகும். *ஊ* என்பது மேல் பின் இதழ் குவி நெட்டுயிராகும். *எ* என்பது மேலிடை முன் இதழ் விரி குற்றுயிரொலியாகும். *ஏ* என்பது மேலிடை முன் இதழ்விரி நெட்டுயிரொலியாகும். *ஒ* என்பது மேலிடை பின் இதழ் குவி குற்றுயிரொலியாகும். ஓ என்பது மேலிடை பின் இதழ் குவி நெட்டுயிரொலியாகும். *அ* என்பது கீழ் நடு இதழ்விரி நெட்டுயிராகும். ஆ என்பது கீழ் நடு இதழ் விரி நெட்டுயிராகும்.

உயிரொலிகளின் ஓசை குரல்வளை மடல்களின் அதிர்வு மற்றும் பேச்சுக்குழாயின் நீளம் இவற்றின் இணைப்பு அடிப்படையில் தீர்மானிக்கப்படுகின்றது. குரல்வளை மடல்களின் அதிர்வு எல்லா உயிரொலிகளுக்கும் குறிகையின் அதிர்வெண் உருப்படுத்தத்தில் பார்க்க இயலும் இயைபுடன் (harmonics) காலப்புலத்தில் நிறைவற்ற காலம்மாறா அலைவடிவைத் (quasi-periodic waveform) தருகின்றது. அதிர்வெண் புலத்தில் இந்த இயைபுகளுக்கு இடையிலான இடைவெளி அல்லது காலப் புலத்தில் ஒரு காலப்பகுதியின் நீளத்தின் பரஸ்பரம் ஓசையின் இசைமை (அல்லது அடிப்படை) நிகழ்வெண் ஆகும். இது பெண் குரலில் பெரும்பான்மையும் அதிகமாக இருக்கும். ஒவ்வொரு சொல்லின் நிறமாலை நிழற்பட திட்டங்களில் குரல் ஒலியன்களின் இயைபு கிடைமட்டமான (குறுகிய பட்டை) (horizontal (narrowband)) அல்லது செங்குத்தான (அகன்ற பட்டை) வரிகளாக/கோடுகளாகக் (vertical (wideband) striations)) காணப்படும் . எல்லா குரலொலிகளும் உயிரொலிகள் அல்ல மற்றும் எல்லா உயிரொலிகளும் குரலொலிகள் என்பதைக் கருத்தில் கொள்ளவும். நிறமாலை நிழற்படத்தில் ஒலிச்செறிவின் இருப்பிட வேறுபாடுகள் (அல்லது தூரங்கள்) இவ்வுயிரொலிகளுக்கு இடையிலான வேறுபாட்டை வெளிப்படுத்தும்.

தமிழில் எல்லா உயிரொலிகளும் குரலொலிகளாகும். அவற்றின் உருவாக்கத்தில் பேச்சுக்குழல் இரண்டு அல்லது மூன்று நன்றாக அடையாளப்படுத்தப்பட்ட ஒலியூக்கங்களை அதாவது உயிரொலிச் செறிவுகளைக் (F1, F2, F3) காட்டும். ஒலிச்செறிவுகள் மனிதப் பேச்சின்

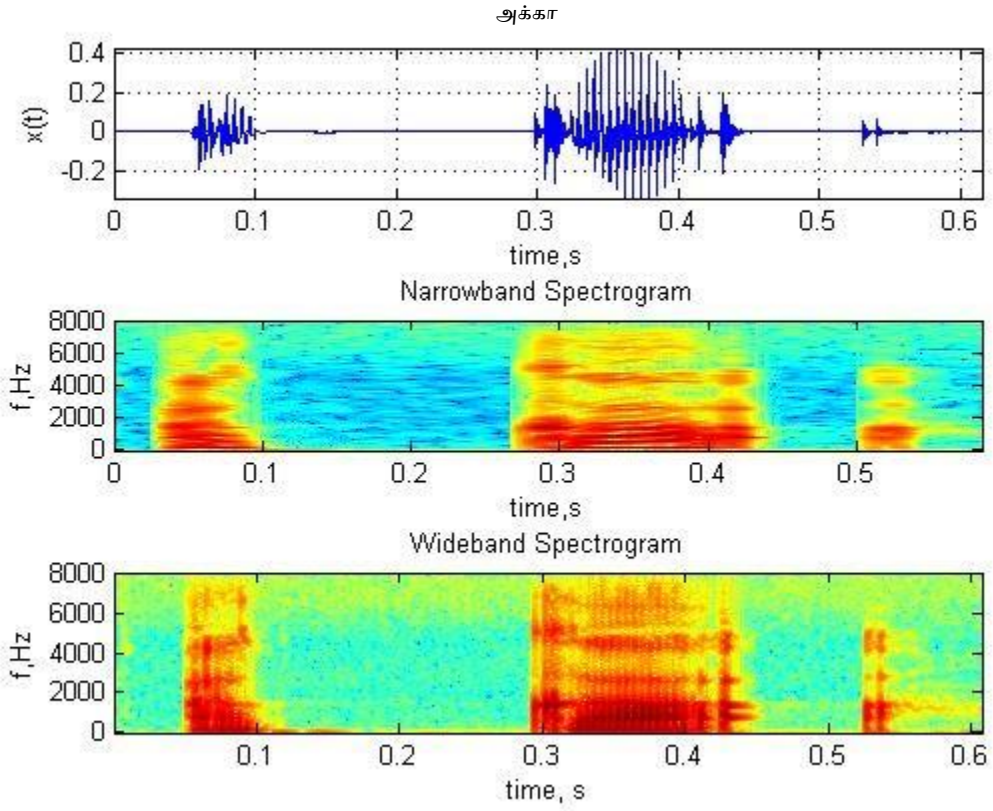
வேறுபடுத்துகின்ற மற்றும் பொருள்வாய்ந்த நிகழ்வெண் கூறுகள் ஆகும். வரையறை விளக்கப்படி உயிரொலிகளுக்கிடையில் வேறுபடுத்த மனிதர்கள் தேவைப்படும் தகவல் உயிரொலிகளின் நிகழ்வெண் பொருளடக்கத்தால் அளவு அடிப்படையில் உருப்படுத்தம் செய்யவியலும். பேச்சில் இந்தச் சிறப்புப்பண்புப் பகுதிகள் தான் கேட்பவர்களுக்கு உயிரொலிகளை அடையாளம் காட்டுகின்றது. பெரும்பாலான இந்த ஒலிச்செறிவுகள் குழாய் மற்றும் அறை ஒலியூக்கத்தால் (tube and chamber resonance) உருவாக்கப்படுகின்றது. குறைந்த அதிர்வெண் கொண்ட ஒலிச்செறிவு F1 எனப்படும்; இரண்டாவது F2 ஆகும்; மூன்றாவது F3 ஆகும். பெரும்பான்மையும் முதல் இரு ஒலிச்செறிவுகளும், F1, F2 என்பன உயிரொலிகளின் மயக்கம் தீர்க்கப் போதுமனதாகும். இந்த இரு ஒலிச்செறிவுகள் திறந்த/மூடிய மற்றும் முன்/பின் பரிமாணங்கள் அடிப்படையில் உயிரொலிகளின் பண்பை தீர்மானிக்கின்றது. இப்பரிமாணங்கள் மரபு ஒலிப்பு விளக்க அடிப்படையில் நாக்கின் இருப்புநிலையைச் சார்ந்து அமையும். முதல் ஒலிச்செறிவு F1 திறந்த உயிரொலிகளுக்கு (அ) உயர்ந்த அதிர்வெண்ணையும் மூடிய உயிரொலிகளுக்கு (இ, உ) குறைந்த அதிர்வெண்ணையும் கொண்டிருக்கும். இரண்டாவது ஒலிச்செறிவு F2 முன்னுயிர்களுக்கு (இ, எ) உயர்ந்த அதிர்வண்ணையும் பின்னுயிர்களுக்கு (உ, ஒ) குறைந்த அதிர்வண்ணையும் கொண்டிருக்கும். உயிரொலிகள் பெரும்பான்மையும் நான்கோ அதற்கு மேலோ வேறுபடுத்தி அறியும் ஒலிச்செறிவுகளைக் கொண்டிருக்கும்; சில வேளைகளில் ஆறுக்கும் மேற்பட்ட ஒலிச்செறிவுகள் இருக்கும். இருப்பினும் உயிரொலியின் பண்பைத் தீர்மானிக்க முதல் இரண்டு ஒலிச்செறிவுகள் முக்கியமாகும்; இது முதல் ஒலிச்செறிவுக்கு எதிரான இரண்டாவது ஒலிச்செறிவின் வரைவு அடிப்படையில் காட்சிப்படுத்தப்படுகின்றது; இருப்பினும் இதழ்குவி போன்ற உயிரொலிகளின் பண்புகளின் சில நோக்குகளைக் கண்டுபிடிக்க போதுமானதல்ல.

உயிரொலிகளைப் பொறுத்தவரையில் ஒலிச்செறிவு அமைப்புகள் அவற்றை வேறுபடுத்தும் ஒலியியக்கக் குறிப்புகளாக உள்ளன. நாம் இந்த ஒழுங்கு முறைக்கு உட்படுகின்ற ஒலிகளை f1-f2 ஒலியியக்கக் குறிப்பு அடிப்படையில் அறிந்து கொள்கிறோம். நெட்டுயிரொலிகளுக்கு குற்றுயிரொலிகளைவிடக் கூடுதல் கால அளவு இருக்கிறது என்பது ஒலியியக்கக் குறிப்புகளால் வெளிப்படும் உண்மை. ஈருயிர்கள் நீண்ட உயிர்களின் வகுப்பில் அடங்கும். இறுதியாக திறந்த உயிரொலொலிகளை மூடிய உயிரொலிகளிடமிருந்து வேறுபடுத்த உதவும் கடுமை ஒலியியக்கக் குறிப்பும் இருக்கிறது. வேறுபடுதலின் மொத்த நீட்சி எல்லை 7dB ஆகும். ஆனால் இந்த ஒலியியக்கக் குறிப்பு உயிரொலிகளை அறிந்து கொள்வதில் பங்கு கொள்கின்றது. இந்த இடங்களில் குரல் ஒலி மற்றும் குரலிலா ஒலி இவற்றின் வேறுபாட்டிற்கான

ஒலியியக்கக் குறிப்பு குரல் ஒலியில், தொண்டை அதிர்வால் ஆக்கிரமிக்கப்பட்ட எல்லையில் தாழ்ந்த நிகழ்வெண்ணின் உள்ளமையும் தொண்டை அதிர்வுடன் தொடர்பு கொண்ட காலநீட்சிமாறாமையின் தொடர்ச்சியும் ஆகும். குரலிலா ஒலிக்கு ஒலியியக்கக் குறிப்பு மேற்சொன்னவை இல்லாதிருப்பதாகும்.

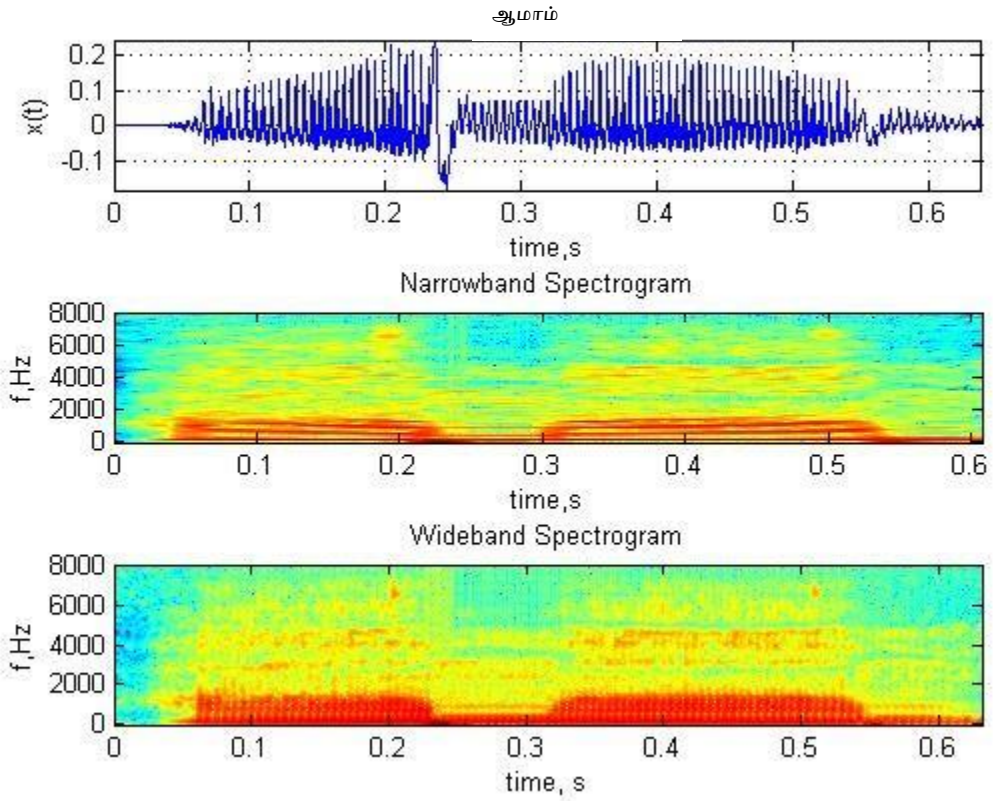
3.4.2.1.1. அ-வின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

அக்கா



3.4.2.1.2. ஆ-வின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு

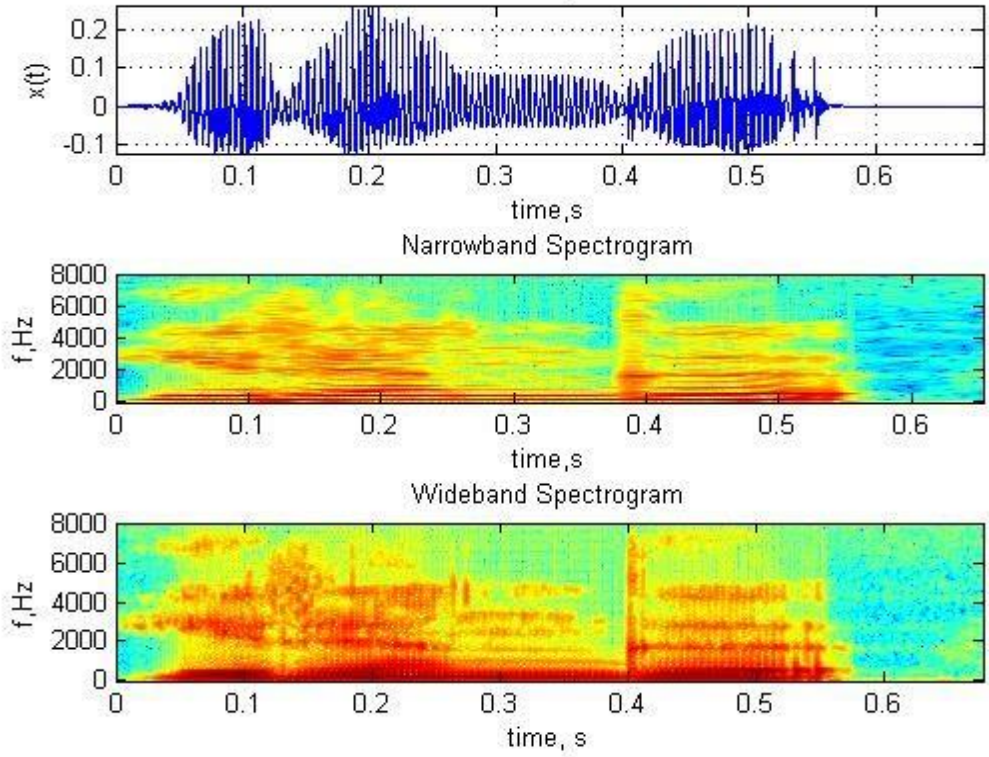
ஆமாம்



3.4.2.1.3. இ-யின் ஒலியியக்கப் பண்புக்கூறுகளை வெளிப்படுத்தும் நிறமாலைவரைவு

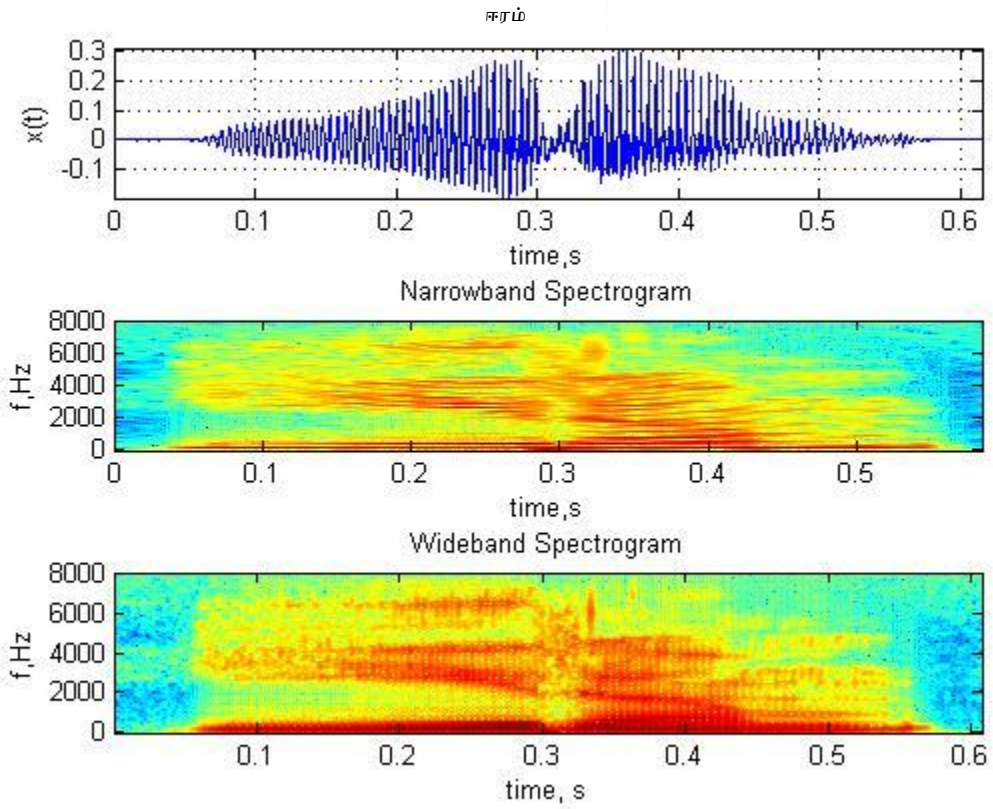
இரங்கு

இரங்கு



3.4.2.1.4. ஈ-யின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு

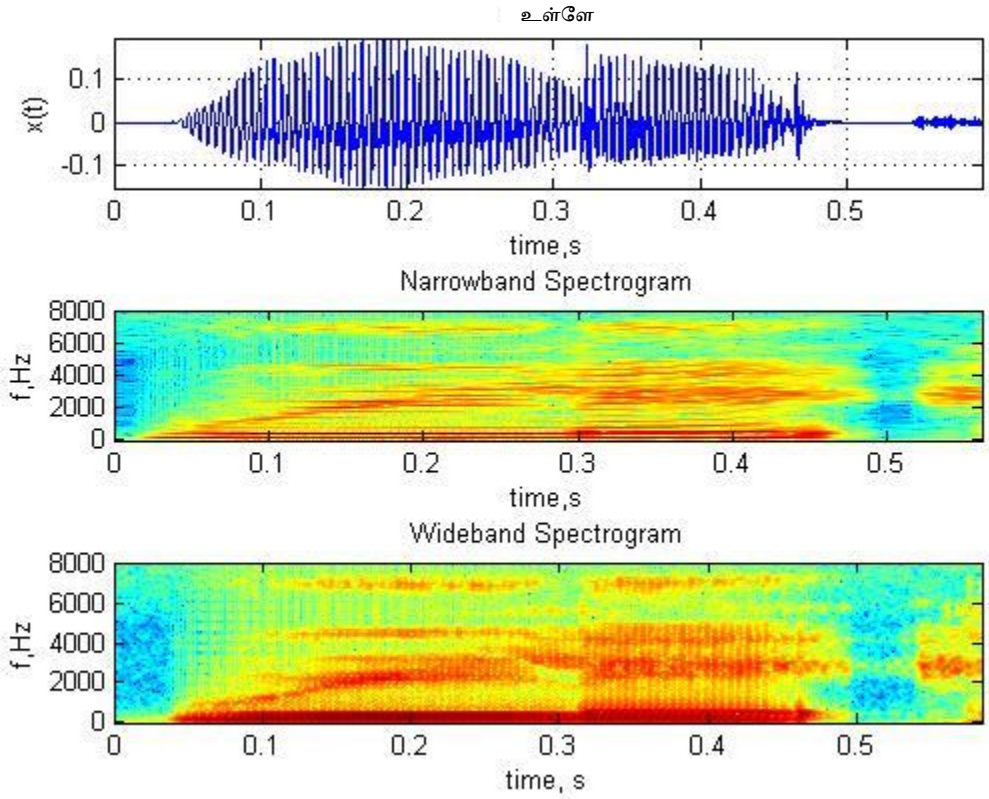
ஈரம்



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

3.4.2.1.5. உ-வின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு

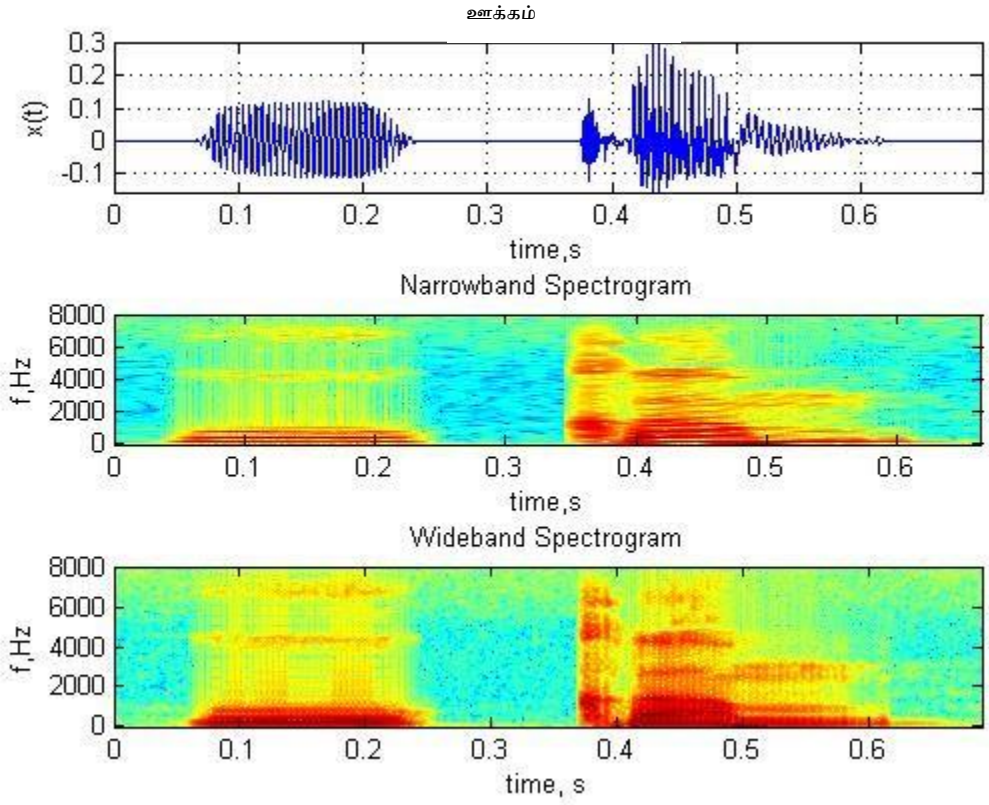
உள்ளே



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

3.4.2.1.6. ஊ-வின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு

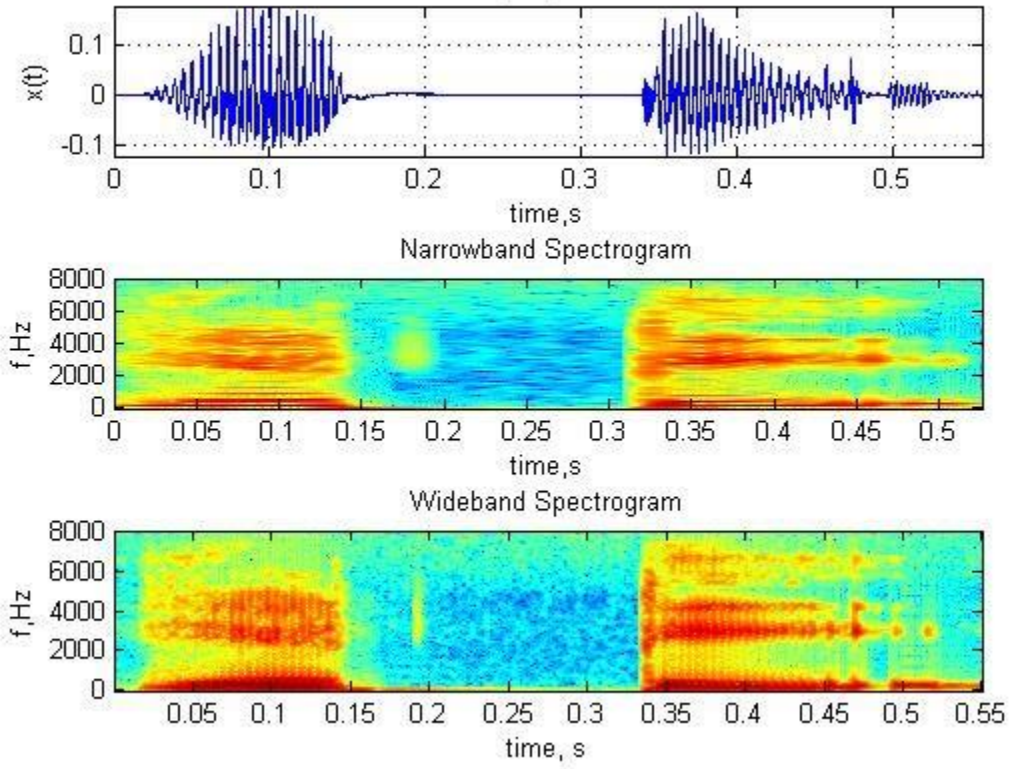
ஊக்கம்



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

3.4.2.1.7. எ-யின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு எத்தி

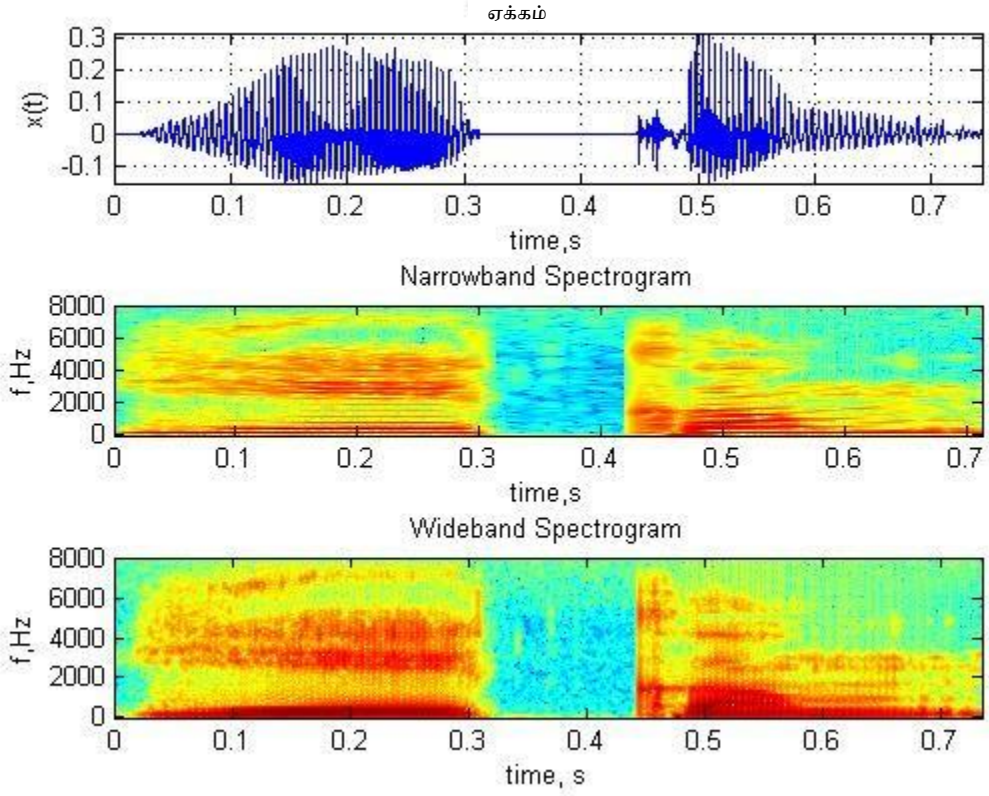
எத்தி



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

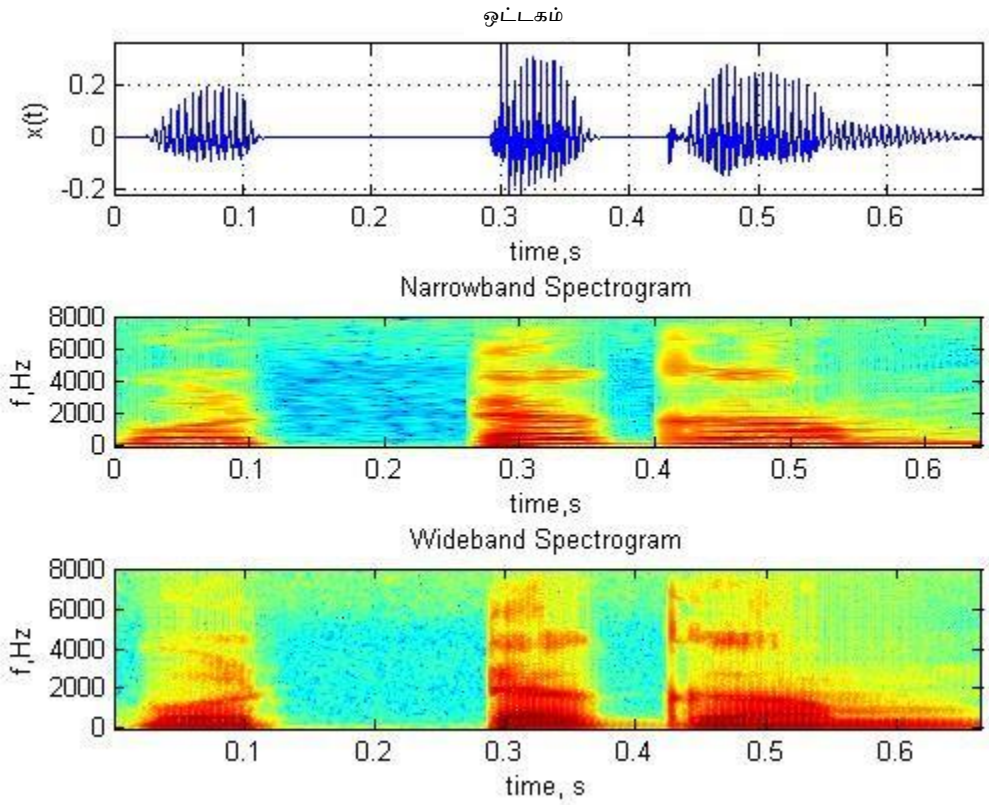
3.4.2.1.8. ஏ-யின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு

ஏக்கம்



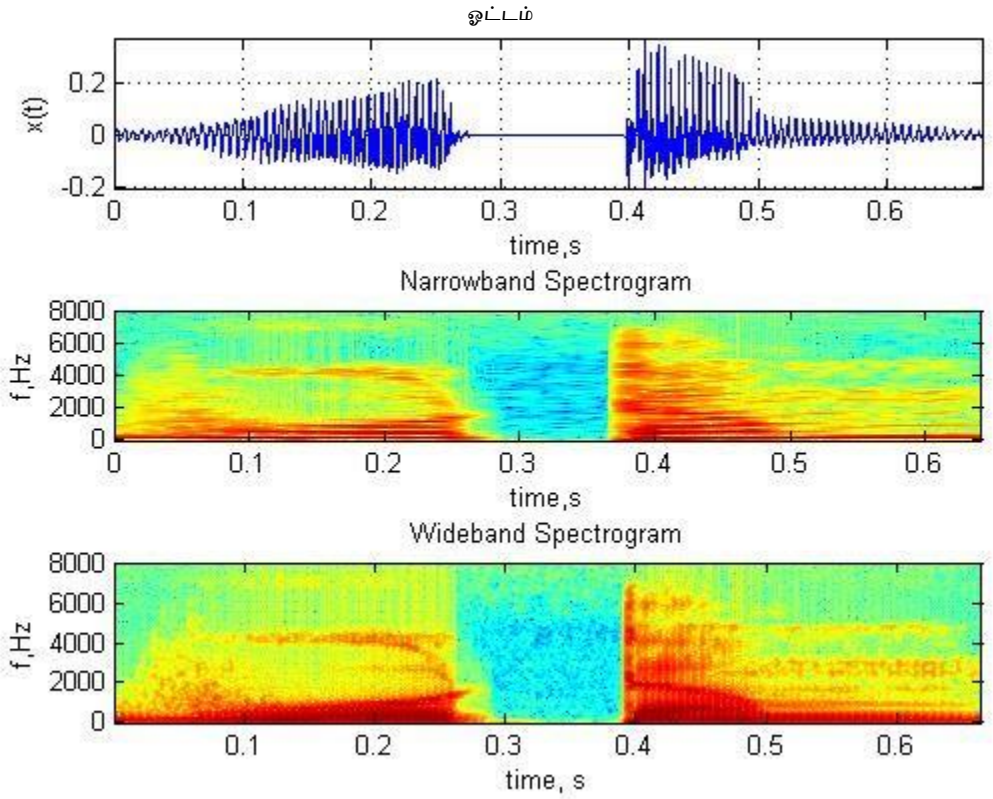
3.4.2.1.9. ஒ-வின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு

ஒட்டகம்



3.4.2.1.10. ஓ-வின் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமாலைவரைவு

ஓட்டம்



3.4.2.2. மெய்யொலிகளின் ஒலியியக்கப் பண்புக்கூறுகள்

தமிழில் உள்ள மெய்யொலிகளை ஒலிப்பியல் அடிப்படையில் அவற்றில் ஒலிப்பிடத்தையும் ஒலிப்புமுறையையும் கருத்தில் கொண்டு பகுத்தாயலாம் என்று முன்னர் பார்த்தோம். ஒலிப்பிடம் அடிப்படையில் *ப்,ம்* என்பவைகளை இதழ் ஒலிகளாகவும் *த், ந்* என்பவைகளை பல்லொலிகளாகவும் *ற், ர், ன்* என்பவைகளை அண்பல்லொலிகளாகவும் *ச், ஞ், ய்* என்பவைகளை இடையண்ண ஒலிகளாகவும் *க், ங்* என்பவைகளை கடையண்ண ஒலிகளாகவும் *ட், ண், ள், ழ்* என்பவைகளை நாமடி ஒலிகளாகவும் பகுத்தாயலாம் எனவும் ஒலிப்பு முறை அடிப்படையில் *க், ச், ட், த், ப்* என்பவைகளை வெடிப்பொலிகளாகவும் *ங், ஞ், ண், ந், ம், ன்* என்பவைகளை மூக்கொலிகளாகவும் *ல், ள், ழ்* என்பவைகளை மருங்கொலிகளாகவும் *ற்* என்பதை ஆடொலியாகவும் பகுத்தாயலாம் என்று முன்னர் பார்த்தோம்.

க், ச், ட், த், ப் வெடிப்பொலிகள் சுற்றி இருக்கிற உயிரொலிகளின் ஒலிச்செறிவுகளின் இருப்பிடத்தை மாற்றும். *ப்* போன்ற ஈரிதழ் ஒலிகள் ஒலிச்செறிவுகளின் இருப்பிடத்தை தாழ்த்தும்; *க்* போன்ற கடையண்ண ஒலிகள் F1, F2 பெரும்பாலும் எப்பொழுதும் சேர்ந்துவருவதை

அனுமதிக்கும். *ற், ர், ன்* போன்ற அண்பல்லொலிகள் எந்த உயிரொலி இருக்கிறது என்பதைப் பொறுத்து அடுத்துவரும் உயிரொலிகளின் ஒலிச்செறிவுகளில் குறைந்த அளவு சீரான மாற்றங்களை ஏற்படுத்தும். உயிரொலி ஒலிச்செறிவு நிகழ்வெண்களில் இந்த மாற்றங்களின் கால ஓட்டம் 'ஒலிச்செறிவு நிலைமாற்றங்கள்' (formant transitions) என்று அழைக்கப்படும்.

வேவ்வேறு இடங்களில் காற்றொழுக்கு அடைக்கப்பட்டு விடுவிக்கப்படுவதால் வெடிப்பொலிகள் உருவாக்கப்படுகின்றன. குரல்வளை மடல்கள் அதிர்வுறுவதால் அவைகள் குரலொலிகளாக வெளிப்படுகின்றன. பேச்சுக்குழாயானது வெவ்வேறு இடங்களில் அடைக்கப்படுவதன் காரணமாக அவற்றின் ஒலியூக்களில் மாற்றங்கள் விளைவிக்கப்படுகின்றன. இது அவற்றின் அதிர்வெண் அமைப்பு வடிவிலும் நிறமாலை நிழற்படங்களிலும் மாற்றங்களை வெளிப்படுத்துகின்றன. வெடிப்பொலிகள் சுற்றியிருக்கும் உயிர்களின் ஒலிச்செறிவுகளின் (F1, F2, F3) இருப்பிடத்தில் மாற்றங்களை விளைவிக்கின்றது. ஈரிதழ் ஒலிகள் (*ப்*) ஒலிச்செறிவுகள் தாழ்வதை விளைவிக்கின்றது. கடையண்ண ஒலிகள் (*க்*) F2-உம் F3-உம் பெரும்பாலும் எப்பொழுதும் சேர்ந்துவருவதை விளைவிக்கின்றது. அண்பல் ஒலிகள் உயிரொலி ஒலிச்செறிவில் அடுத்துவரும் உயிரொலிகலின் ஒலிச்செறிவுகளில் குறைந்த அளவிலேயே சீரான மாற்றங்களை ஏற்படுத்தும்; இது எந்த உயிர்கள் வருகின்றன என்பதைப் பொறுத்து அமையும்.

வெடிப்பொலிகளின் ஒலிப்பின் போது காற்றொழுக்கானது தடைபடுவதால் அதாவது பேச்சுக்குழல் சில கணங்கள் மூடப்படுவதால் அது மௌனத்தை உருவாக்கும். இது நிறமாலை நிழற்படத்தில் காலியிடமாகத்தெரியும். ஏனென்றால் இங்கு பதிவு செய்வதற்கு ஒலி சக்தி இல்லை. அடைப்பு விடுவிக்கப் படும்போது ஓசையின் குறுகிய வெடிப்பு வருவதால் அது கால அச்சில் மிகக் குறுகிய ஒழுங்கற்ற சுவட்டை உருவாக்கும். வெடிப்பொலிகளின் ஒலிப்பிடம் அடிப்படையில் மௌனம் அல்லது அடைப்பின் கால அளவு வேறுபடும். இது நிறமாலை நிழற்படத்தில் பதிவு செய்யப்படும். சில வெடிப்பொலிகளின் கால அளவு மிகக் குறுகி இருக்கும். *த்* உடன் தொடர்புடைய கால அளவு கிட்டத்தட்ட 20 மில்லி செகண்ட் ஆகும்; க் உடன் தொடர்புடைய கால அளவு கிட்டத்தட்ட 115 மில்லி செகண்ட் ஆகும்.

வாய் பாதை அடைபட்டு கடையண்ணம் கீழிறங்கும் பொழுது காற்றொழுக்கு மூக்குக் குழாய் வழி திசை திருப்பப்படுகையில் மூகொலிகள் விளைவிக்கப்படும். அடைப்பைப் பொறுத்தமட்டில் *ங் ,ஞ், ண், ந், ம்,ன்* என்ற மூக்கொலிகள் முறையே *க், ச், ட், த், ப், ற்* என்ற வெடிப்பொலிகளுடன் பொருத்தமுடையன. ஆனால் பேச்சுக்குழாயின் நீளமும் பரிமாண அளவும்

காற்று வழியும் வேறுபடுகின்றன. இதன் காரணமாக மூக்கொலிகள் தொடர்புடைய வெடிப்பொலிகளிலிருந்து ஒலியியக்க அடிப்படையில் வேறுபடுவதை உணரலாம்.

ஒலியிக்க இயலில் ஒலியியக்கக் குறிப்புகளை அறியும் ஆய்வு மிகவும் பயன் உள்ளதாகும். பேச்சொலி பற்றிய பல செய்திகளை ஒலியிக்கக் குறிப்புகள் நமக்குத் தெரிவிக்கின்றன. செற்கள் ஒலியன்களின் கோர்வையால் ஆனதாக அறியப்படுகின்றது. பேச்சை ஏற்கும் முதல் செயல்முறை ஒரு குறிப்பிட்ட ஒலியனின் வகைப்பாட்டை உருப்படுத்தம் செய்வதாக உள்வரும் ஒலிகளைக் கண்டுப்பிடிப்பதாகும். ஒரு மொழியின் முழு ஒலியனியல் ஒழுங்குமுறை உறவுகளின் தன்னிறைவான ஒழுங்குமுறையாக நடைபெறுகின்றது. அதாவது ஒரு ஒலியைப் புரிந்து கொள்ளுதல் அதை ஒரு ஒலியன் வகைப்பாட்டை சார்ந்த ஒன்றாகக் கண்டு கொள்வது மட்டுமல்ல, அது பிற ஒலியன் வகைப்பாட்டைச் சார்ந்ததல்ல என்பதைத் தீர்மானிப்பது ஆகும். ஒரு ஒலி ஒரு குறிப்பிட்ட வகைப்பாட்டைச் சார்ந்தது என்பது அவைகளுக்கிடையில் பல பொதுவான ஒலியியக்கக் குறிப்புகளை வெளிப்படுத்தும். மட்டுமன்றி ஒரு வகைப்பாட்டைச் சார்ந்த ஒலிகள் தம்மில் எவ்வாறு வேற்படுகின்றன என்பதையும் ஒலியியக்கக் குறிப்புகள் வெளிப்படுத்தும்.

பேச்சின் அமைப்பொழுங்கைக் காட்சியாக மாற்றுவதால் அதன் மூலம் ஒலிகளை அறிந்து கொள்வதற்கான முக்கியப் பண்புக்கூறுகளைக் கண்டுகொள்ள இயலும் என்று நம்பப்படுகின்றது. ஒலிவர்ணனைப்பட ஆய்வு குறிப்பிட்ட ஒலியில் இருக்கிற சில பண்புக்கூறுகளைக் கூறுகின்றது. மேலும் அந்த பண்புக்கூறு ஒரே விதமான ஒலிகளில் மீண்டும் மீண்டும் வருவதையும் கூறுகின்றது. இருப்பினும் இந்த பண்புக்கூறைத்தான் கேட்பவர்கள் உண்மையிலேயே ஒலியியக்கக் குறிப்பாகப் பயன்படுத்துகின்றனர் என்பதைத் தீர்மானமாக ஒலிவர்ணனைக் கூறவில்லை.

பெரும்பாலும் ஒலியியக்க வர்ணனைப்படங்களை ஆய்ந்து நாம் வேறுப்படுத்துகிற ஒலியியக்கக் குறிப்புகளைக் கண்டுபிடிக்கலாம். எடுத்துக்காட்டாக, ஒரு வெடிப்பு மெய்யொலி (plosive consonant) அதன் அடைப்பு நிலையிலிருந்து விடுவிக்கப்படும் போது அதனால் விளைகிற இரைச்சலின் வெடிப்புக்கு நிகழ்வெண் அளவையில் ஏதாவது ஒரு இடத்தில் சக்தியின் முக்கிய உச்சம் இருக்கும். இந்த உச்சத்தின் இடம் ஒலியின் ஒலிப்பிடம் அடிப்படையில் மிகச்சீராக மாறுபடும். அதாவது ஈரிதழ் அல்லது அண்பல் ஒலி அல்லது கடையண்ண ஒலி என்பதன் அடிப்படையில் மாறுபடும். ஒரு பேச்சு உருவாக்கியின் உதவியால் நாம் அசைகளின் முழு வரிசைகளையும் நிறுவ இயலும்; இதில் தொடக்க இரைச்சல் வெடிப்பு குறைந்த அதிர்வெணிலிருந்து வேறுபடச் செய்யலாம். நம்முடைய ஆய்வுத்தரவு உருவாக்கப்பட்ட ஒலிகளின் மாறுபாடுகளின் எல்லை /b/, /d/, /g/ என்ற மூன்று வகுப்புகளையும் உட்படுத்துவதாக எதிர்பார்க்கலாம் என்று கூறுகின்றது.

தமிழில் ஒலியனியல் அடிப்படையில் அல்லது எழுத்தியல் அடிபடையில் வெடிப்பொலிகளுக்கு/அடைப்பொலிகளுக்கு இடையில் வேறுபாடுகள் இல்லை என்றாலும், ஒலியியல் நிலையில் வேறுபடுகின்ற ஒலிகள் எழுத்துக்களின் அல்லது ஒலியன்களின் மாற்றொலிகளாக வருகின்றன. அவை வரும் இடங்கள் முரண்படாமல் துணைநிலை வருகைமுறையில் வரும். இவ்வாறு *க், ச், ட், த், ப், ற்* என்ற எழுதுக்களுக்கு அல்லது ஒலியன்களுக்கு குரலொலி-குரலிலா ஒலி என்பதன் அடிப்படையில் இரண்டிரண்டு மாற்றொலிகள் உள்ளன. ஒலிப்பு முறை மற்றும் ஒலிப்பு இடம் என்பன வேறுபாட்டாலும் குரலொலி குரலிலா ஒலி வேறுபாட்டிற்கான ஒலியியக்கக் குறிப்பு எல்லா இணைகளுக்கும் பொதுவானதாகும். குரலொலி மற்றும் குரலிலா ஒலி என்பன ஒலியின் ஒலிப்பின் போது குரல்வளை மடல்களின் அதிர்வையோ அதிர்வு இல்லாமையோ குறித்து நிற்கும். சில சூழல்களில் இந்த முரண்பாடு தக்க வைக்கப்படுகின்றது. பல தமிழ் பேசுபவர்களுக்கு மெய்யொலிகள் உயிரிடையில் வரும்போதும் மூக்கொலிகளுக்கு முன்னர் வரும் போது தான் இது சாத்தியமானது. *ற்* என்ற எழுத்து உயிரொலிகளுக்கிடையில் அதிரொலியாக வரும்.

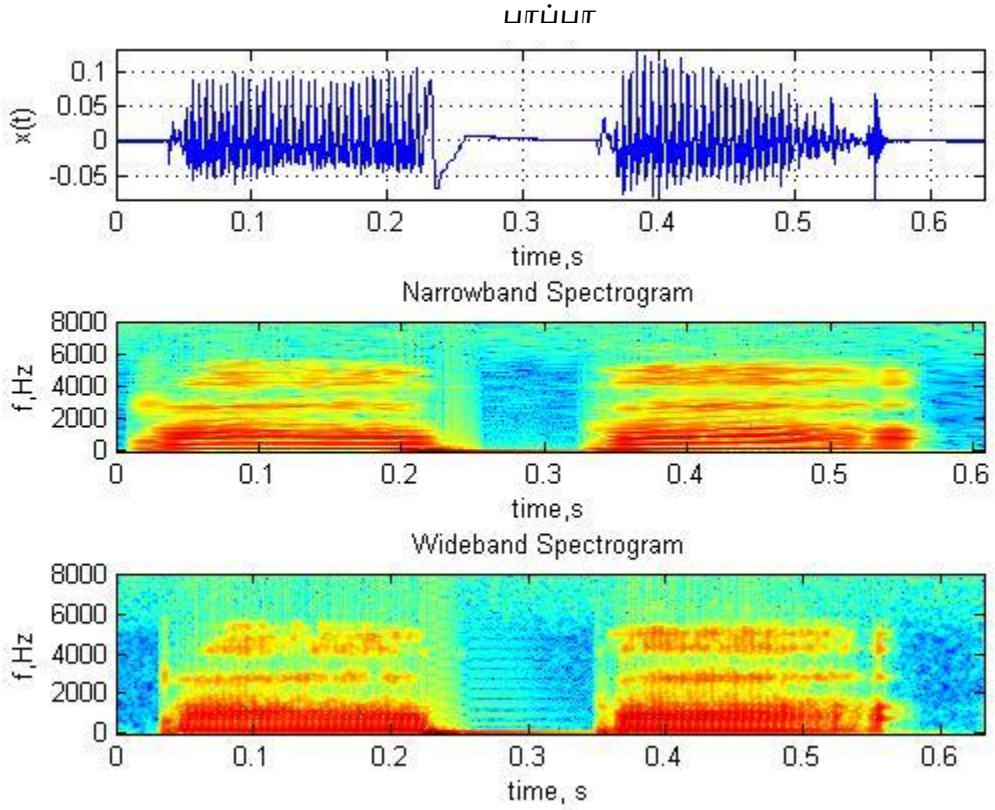
வெடிப்பொலிகள் அவைகளைத் தொடர்ந்து உயிர்வருகிறதா மெய்வருகிறதா என்பதன் அடிப்படையிலும் வேறுபடும். ஒலிப்பின் போது ஒரு கணத்தில் வெடிப்பொலியின் விடுதலையை அடையாளப்படுத்தும் போது ஓசையின் வெடிப்பு (burst of noise) ஏற்படும் அதே கணத்தில் குரல்வளை மடல்களின் அதிர்வுகள் கால நீட்சி மாறாத ஒலிகளுக்காகத் தொடங்கும். குரல் குரலிலா ஒலிகளின் வேறுபாட்டிற்கான மற்றொரு ஒலியியக்கக் குறிப்பு பின்வருமாறு அமையும். மெய்யொலியின் உருவாக்கலின் மீதான எந்த இரைச்சலும் (noise) குரல் ஒலியைவிட குரலிலா ஒலியில் கூடுதல் கருமையைக் கொண்டிருக்கும். எவ்வளவு குறைவாக சக்தி பயன்படுத்தப்படுகின்றதோ அதைப்பொறுத்து ஓசை உற்பத்தியில் செலவழிக்கப்படும். இதன் காரணமாக குரலிலா ஒலியில் ஓசை உயர்ந்த கடுமையில் இருப்பதுடன் நீண்ட நேரம் நிகழும்.

மூக்கொலிகள், மருங்கொலிகள் மற்றும் அரையுயிர்கள் முந்தைய வகுப்புகளிலிருந்து இரைச்சலின் (noise) இல்லாமையாலும் தொடர்ந்த சுரத்தின் உள்ளமையாலும் வேறுபடுத்தப்படும். மூக்கொலிகள் யாவும் தாழ்ந்த நிகழ்வெண் ஒலியூகத்தாலும் இந்த நிகழ்வெண் பட்டைக்கும் 2000 hz இடத்தில் தொடங்கும் பட்டைக்கும் இடையில் சக்தியின் இல்லாமையாலும் ஒலியியக்கக் குறிப்பு செய்யப்பட்டிருக்கும். மருங்கொலிகளும் அரையுயிர்களும் உயிரொலி போன்ற ஒலிச்செறிவுகளால் அடையாளப்படுத்தப்பட்டிருக்கும். இந்த ஒலிச்செறிவுகள் ஒலிச்செறிவு நிகழ்வெண்ணில் ஒத்தறி அடிப்படையில் மெதுவான வேறுபாடுகளைக் கொண்டிருக்கும். இருப்பினும் கடப்புகள்/நிலைமாற்றங்கள் (transitions) அரையுயிர்களைவிட மருங்கொலிகளில் கூடுதல் வேகமாக இருக்கும்.

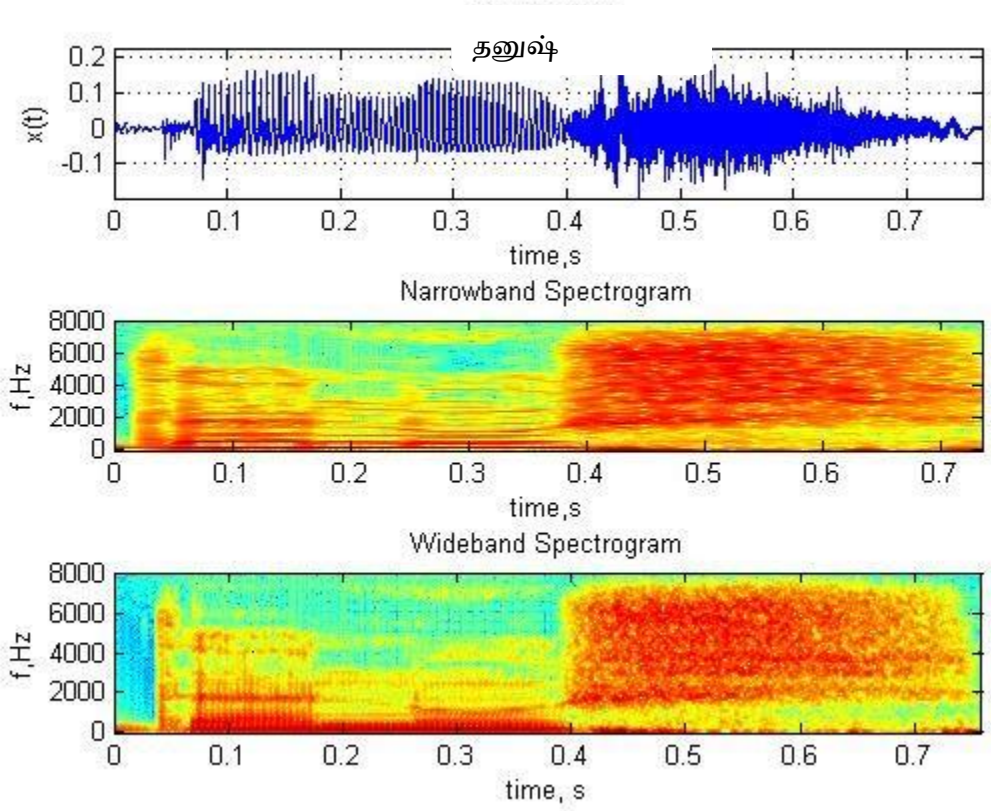
ஒலியிக்கக் குறிப்புகள் ஒலிகள் ஒலிக்கும் இடம் அடிப்படையிலும் வேறுபடும். இதழ் ஒலிகள் (*ப், ம்*), பல்லொலிகள் (*த், ந்*), அண்பல்லொலிகள் (*ற், ர், ன்*), இடையண்ண ஒலிகள் (*ச், ஞ், ய்*), கடையண்ண ஒலிகள் (*க், ங்*), நாமடி ஒலிகள் (*ட், ண், ள், ழ்*) என்பன ஒலிக்கும் இடம் அடிப்படியில் வேறுபடும். இவ்வேறுபாடுகளை உணர்த்தும் ஒலியியக்கக் குறிப்புகளையும் நாம் கண்டுபிடிக்க இயலும்.

ஒலிப்பதில் அல்லது கேட்டுணர்வதில் சிக்கலை விளைவிக்கும் *ல், ள், ழ்* என்ற ஒலிகள் ஒலிப்பிடம் அடிப்படையிலும் ஒலிப்பு முறை அடிப்படையிலும் வேறுபடுவதை நாம் அறிவோம். இவற்றை வேறுபடுத்தும் ஒலியியக்கக் குறிப்புகளை நம்மால் நிறமாலை நிழற்படங்களிலிருந்து கண்டறியலாம். இது போல் *ன், ண்* என்பனவும் வேறுபடுவதை நிறமாலை நிழல் படங்களிலிருந்து கண்டறியலாம். அதிர்வொலியான *ற்* என்பதையும் வருடொலியான *ர்* என்பதையும் வேறுபடுத்தும் ஒலியியக்கக் குறிப்புகளையும் நாம் நிறமாலை நிழற்படங்கள் மூலம் அறிந்து கொள்ளலாம்.

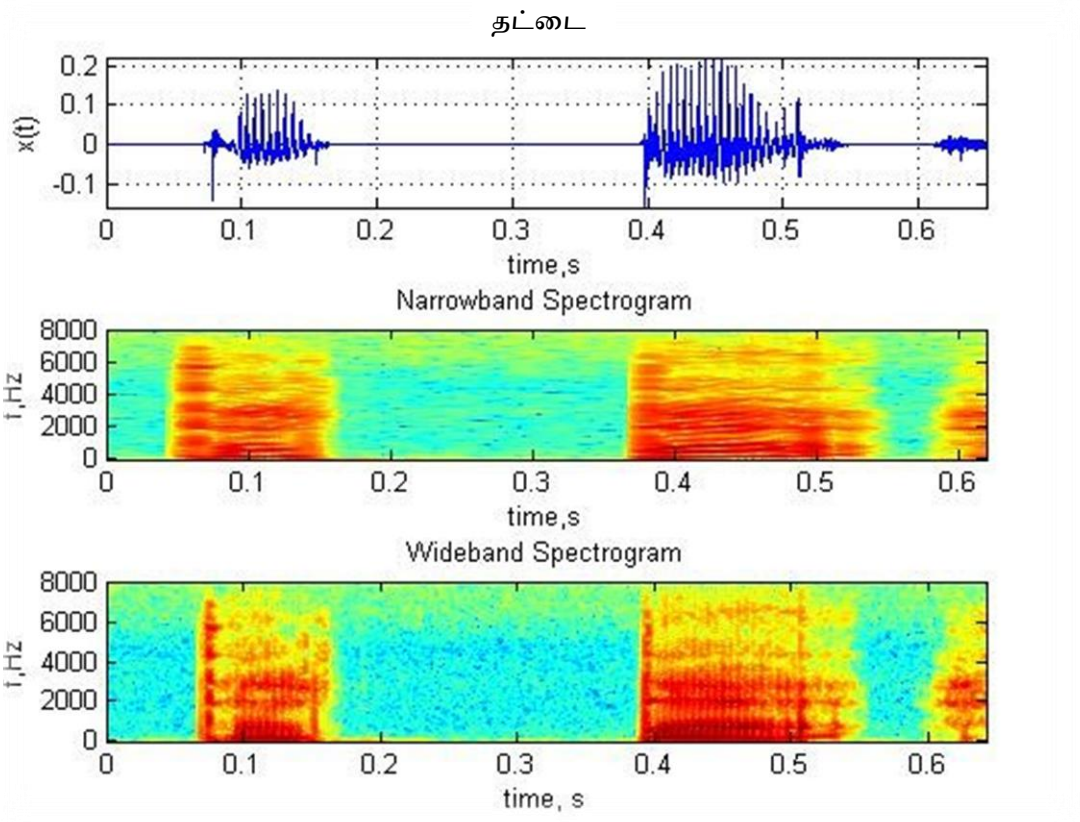
3.4.2.2.1. ப்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு



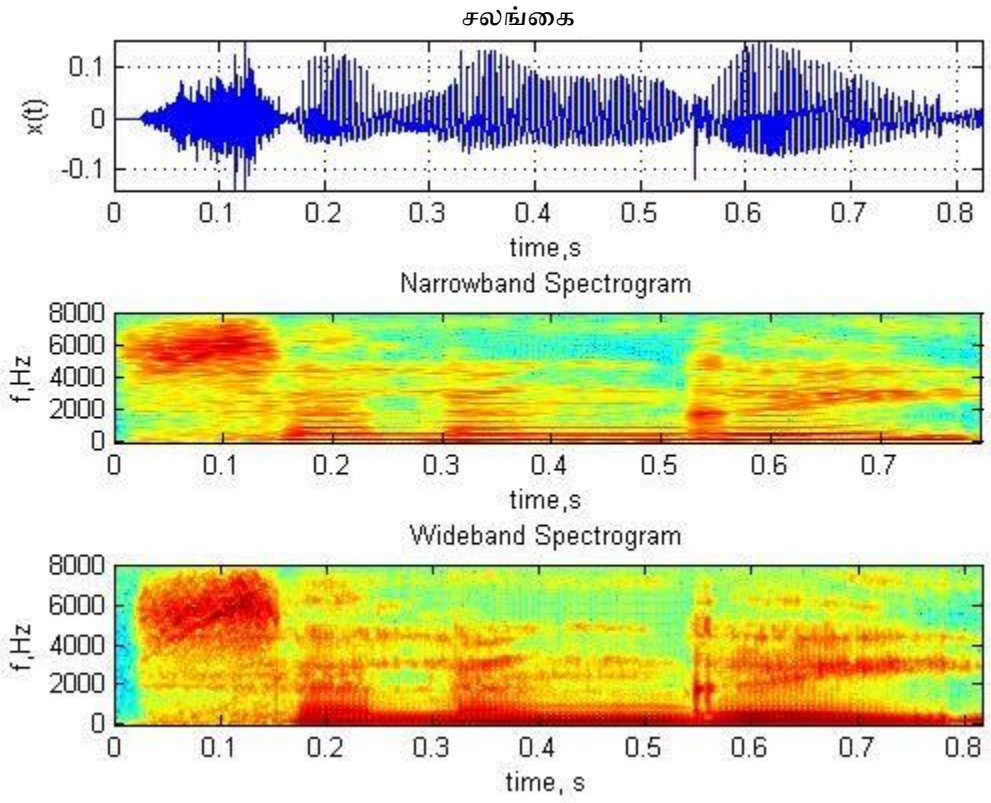
3.4.2.2.2. த்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு



3.4.2.2.3. ட்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

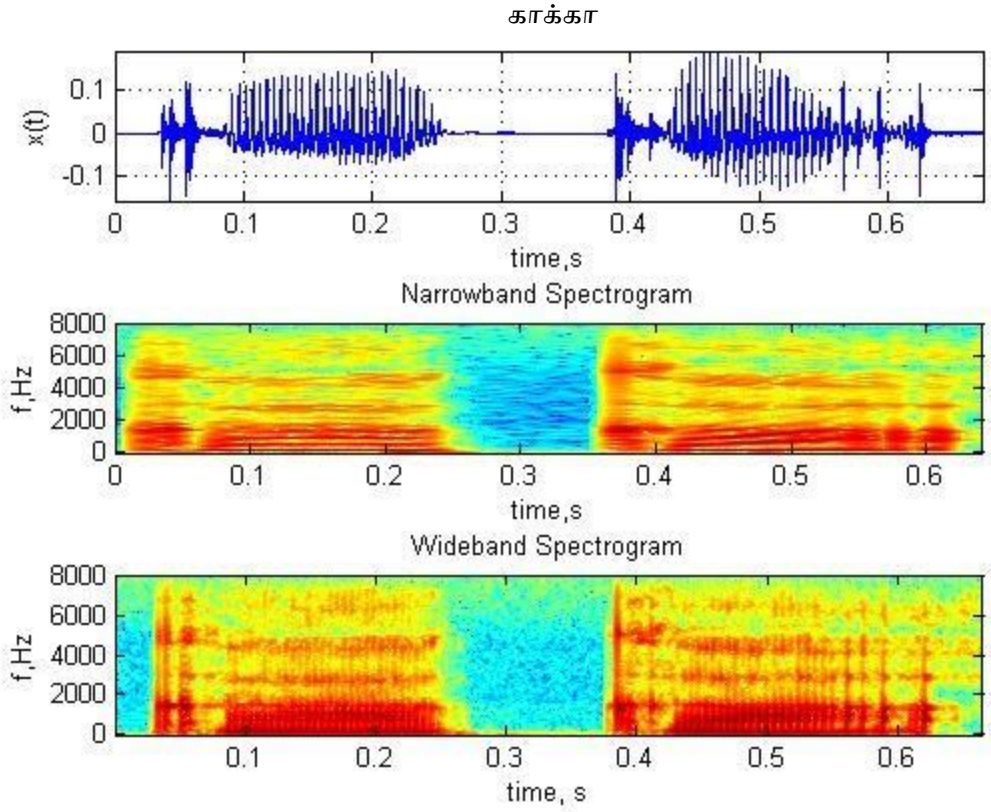


3.4.2.2.4. ச்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

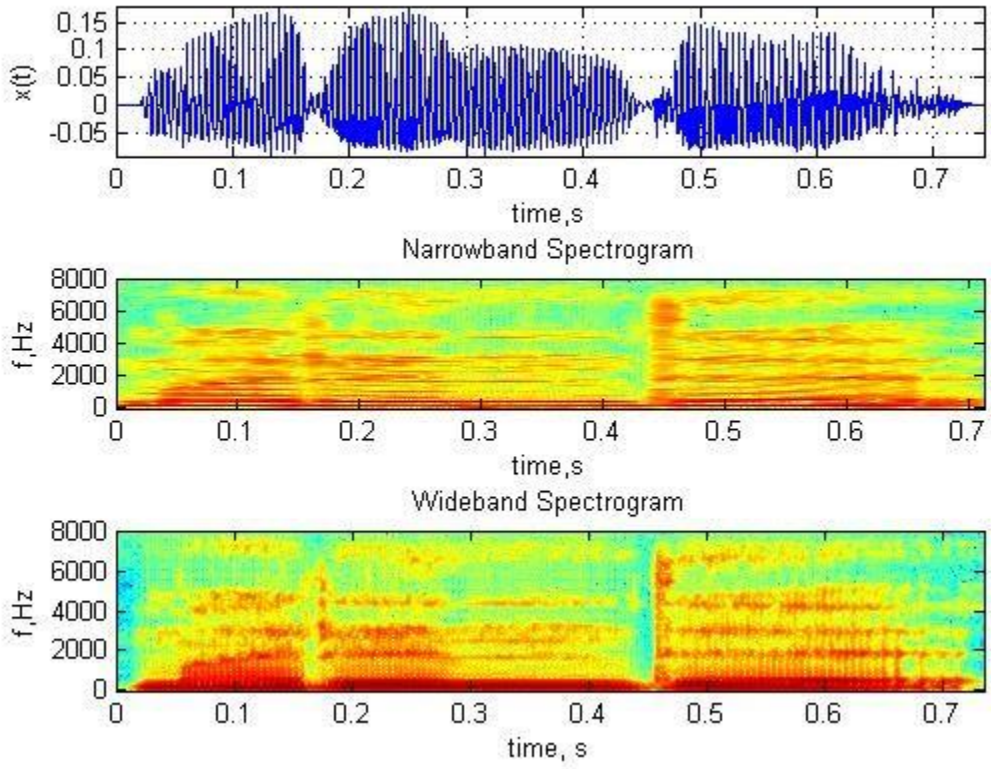
3.4.2.2.5. க்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு



3.4.2.2.6. ம்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

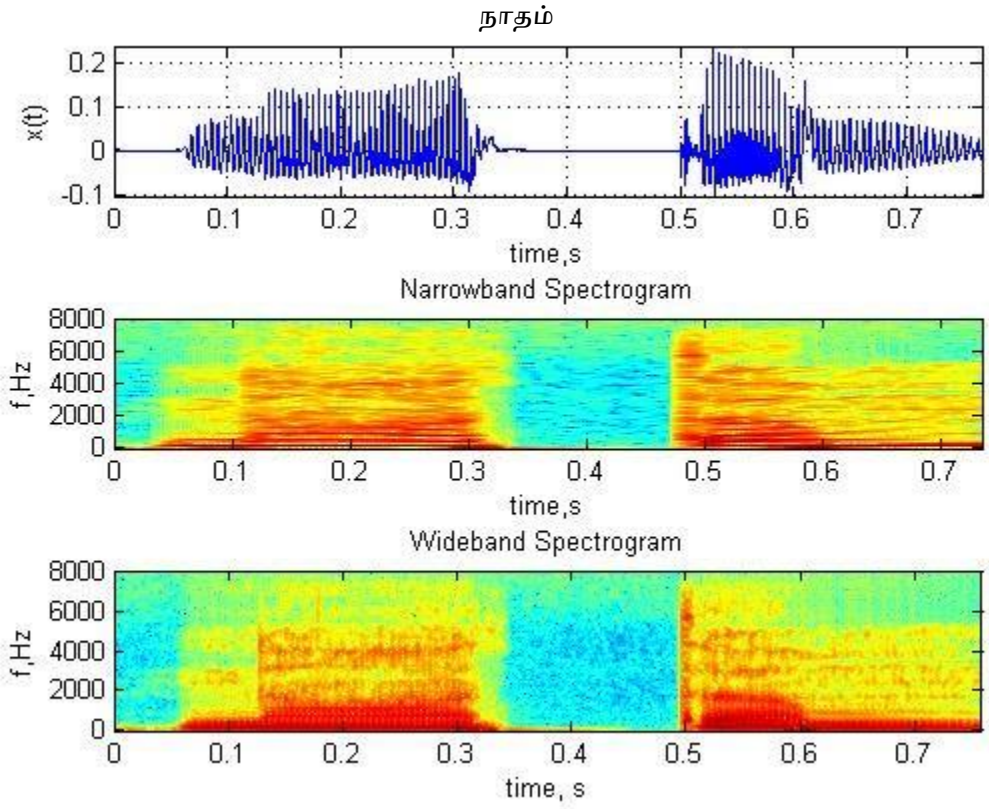
மருந்து

மருந்து



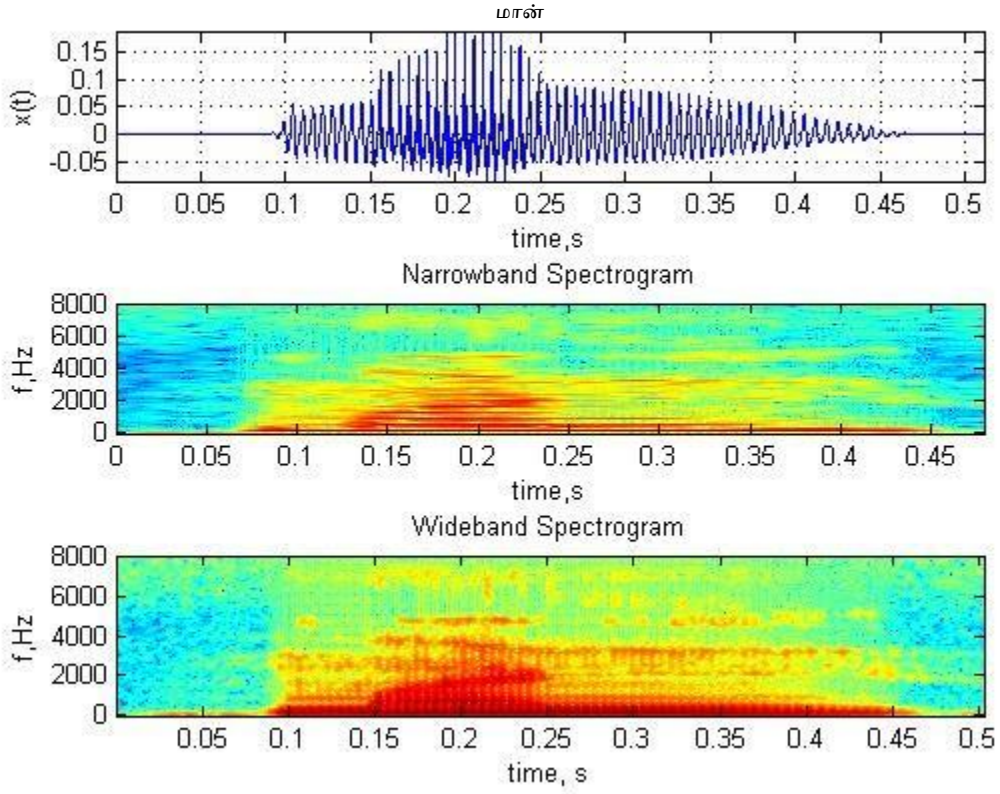
3.4.2.2.7. ந்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

நாதம்



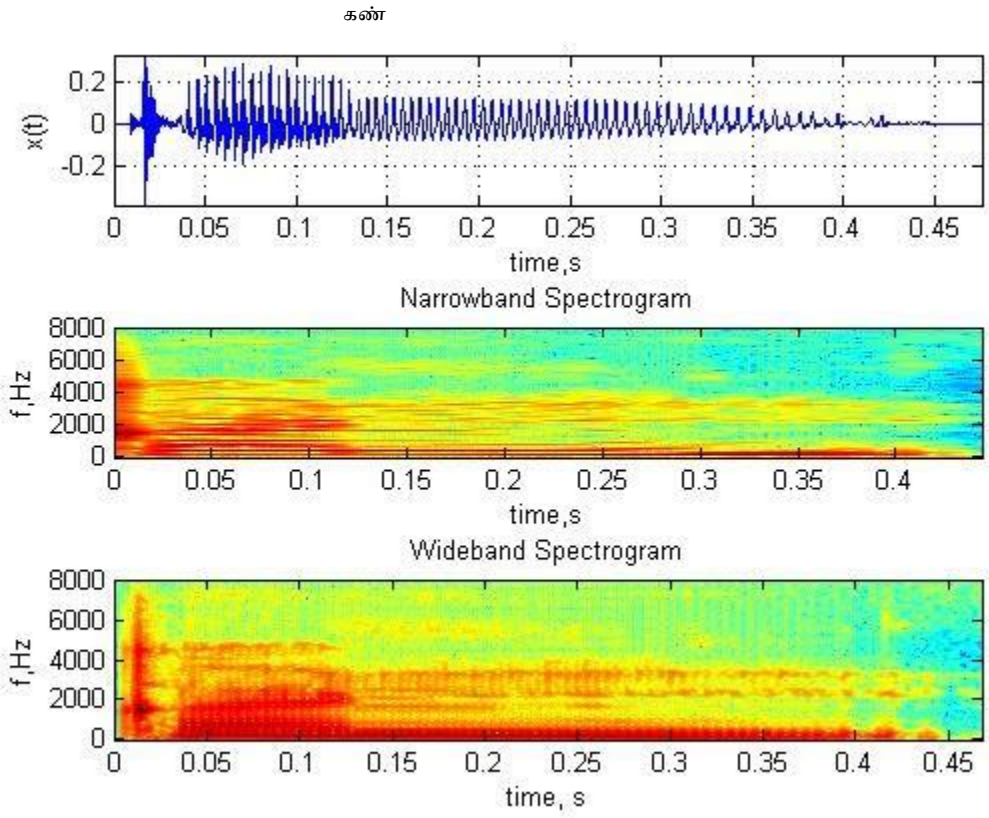
3.4.2.2.8. ன் – ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

மான்



3.4.2.2.9. ண்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

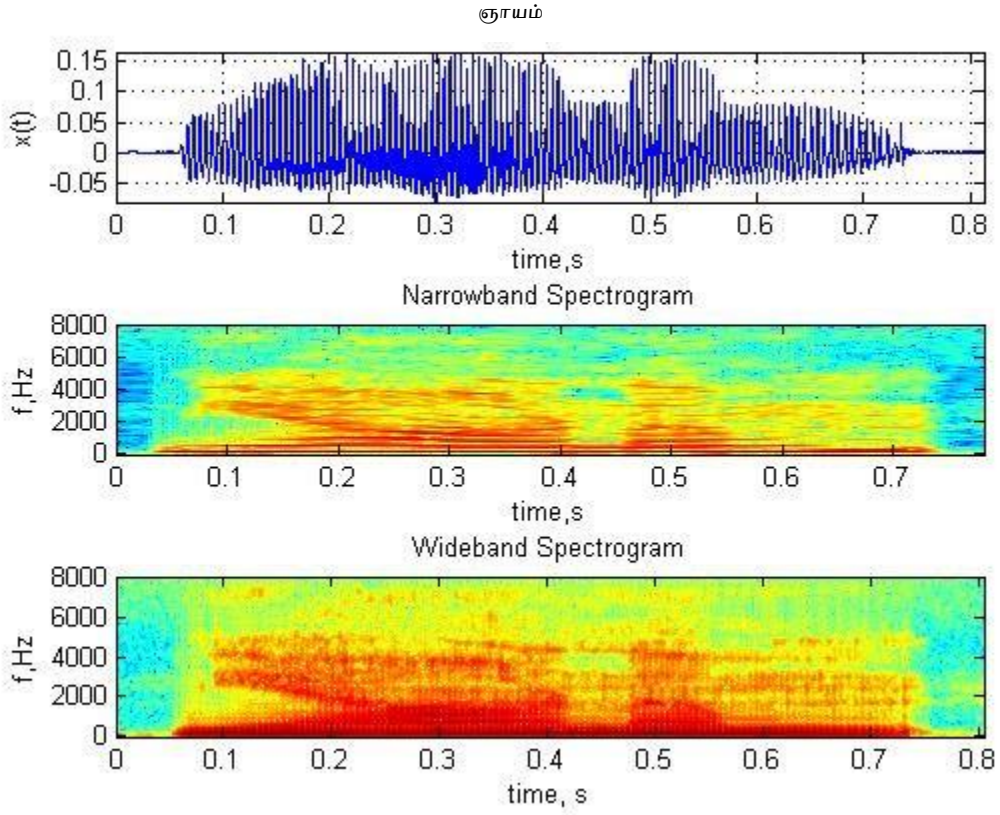
கண்



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

3.4.2.2.10. ஞ்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

ஞாயம்

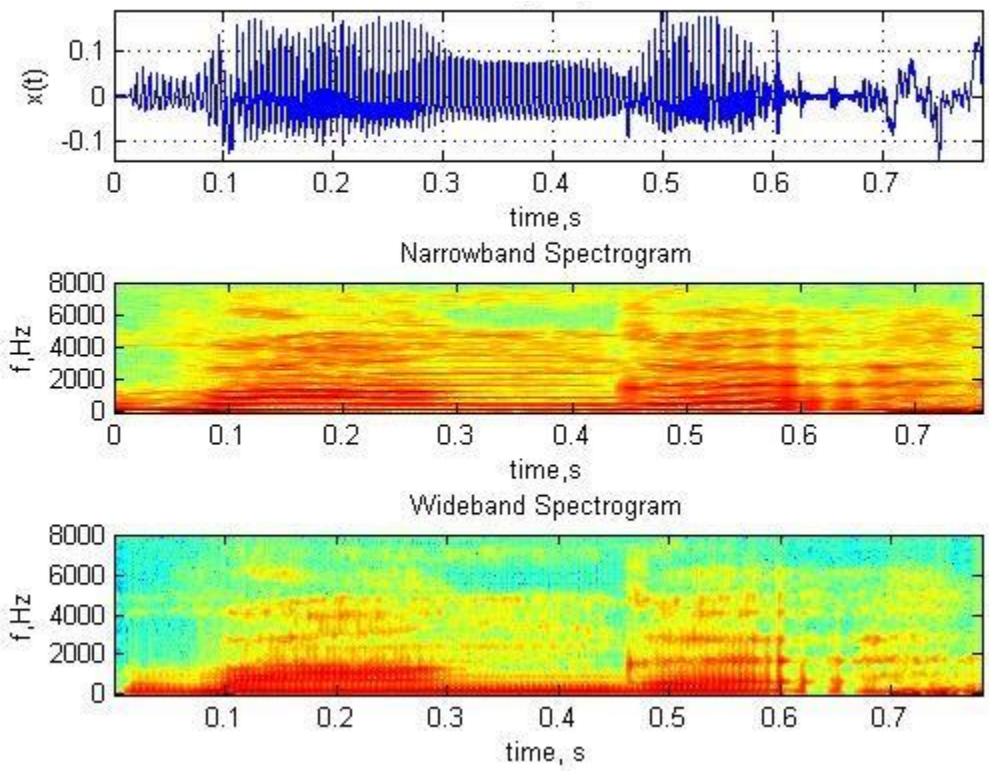


Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

3.4.2.2.11. ங்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

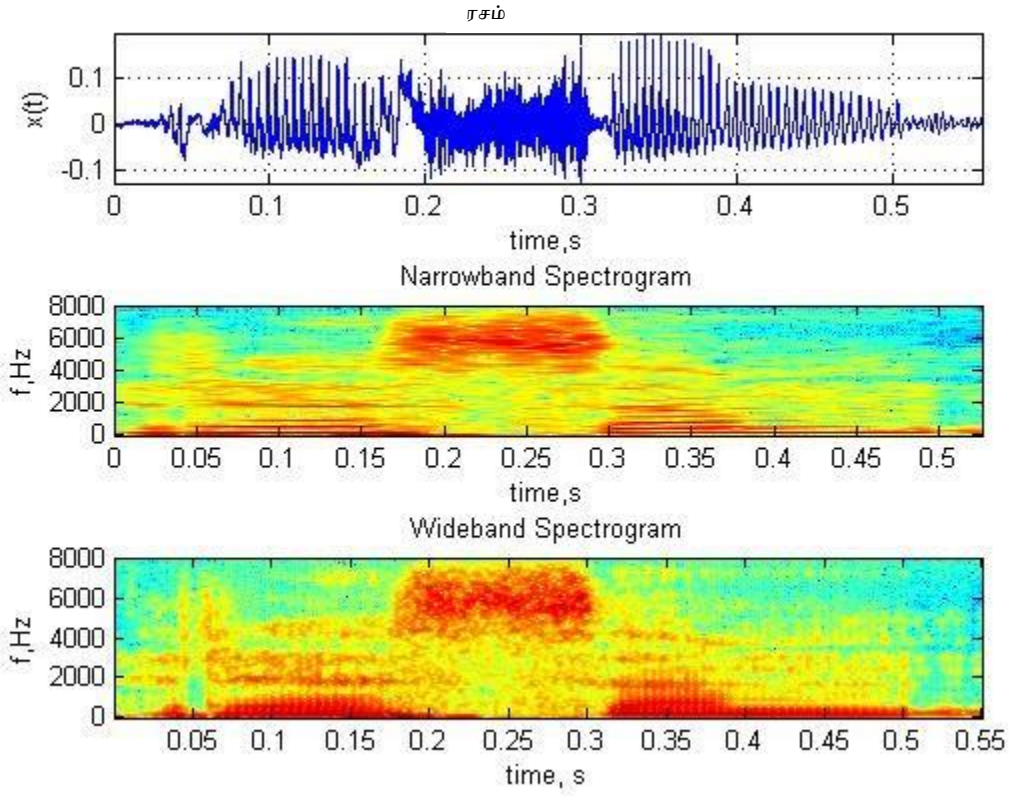
வாங்க

வாங்க



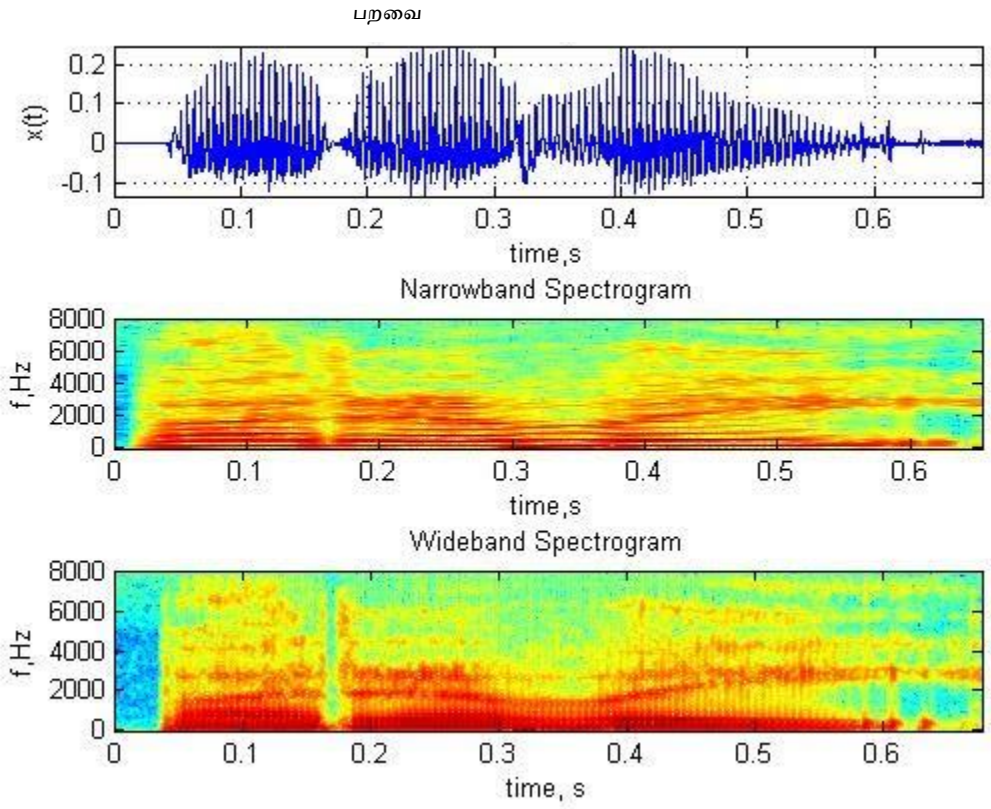
3.4.2.2.12. ர்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

ரசம்



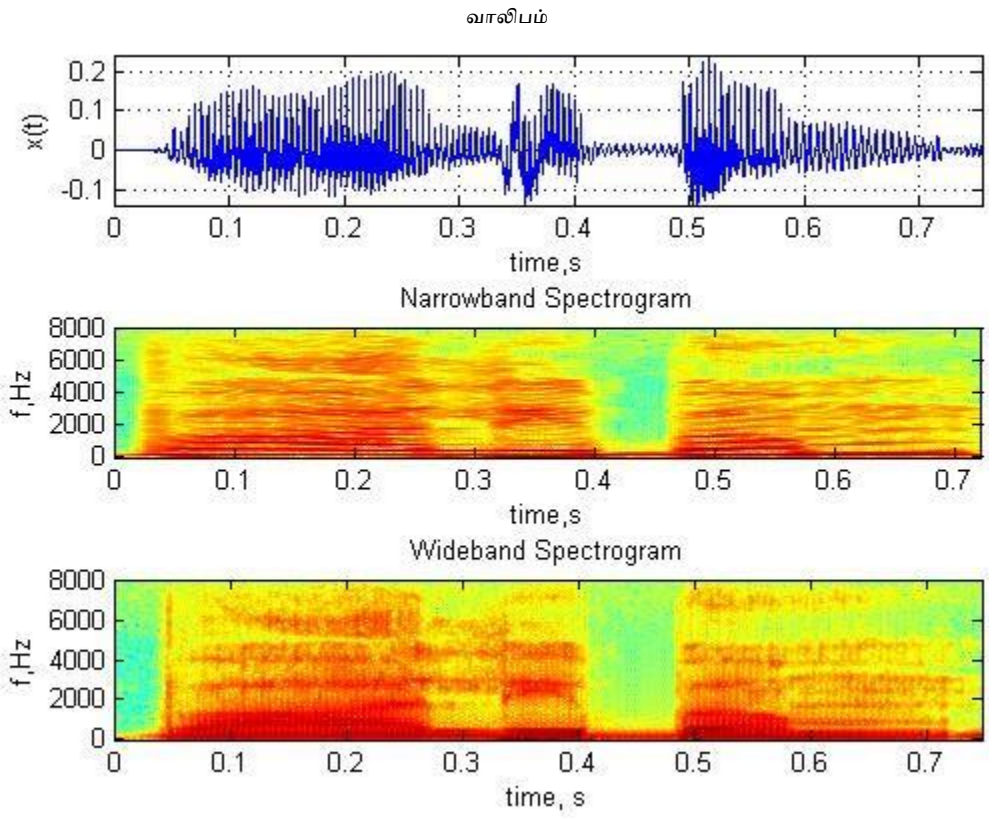
3.4.2.2.13. ற்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

பறவை



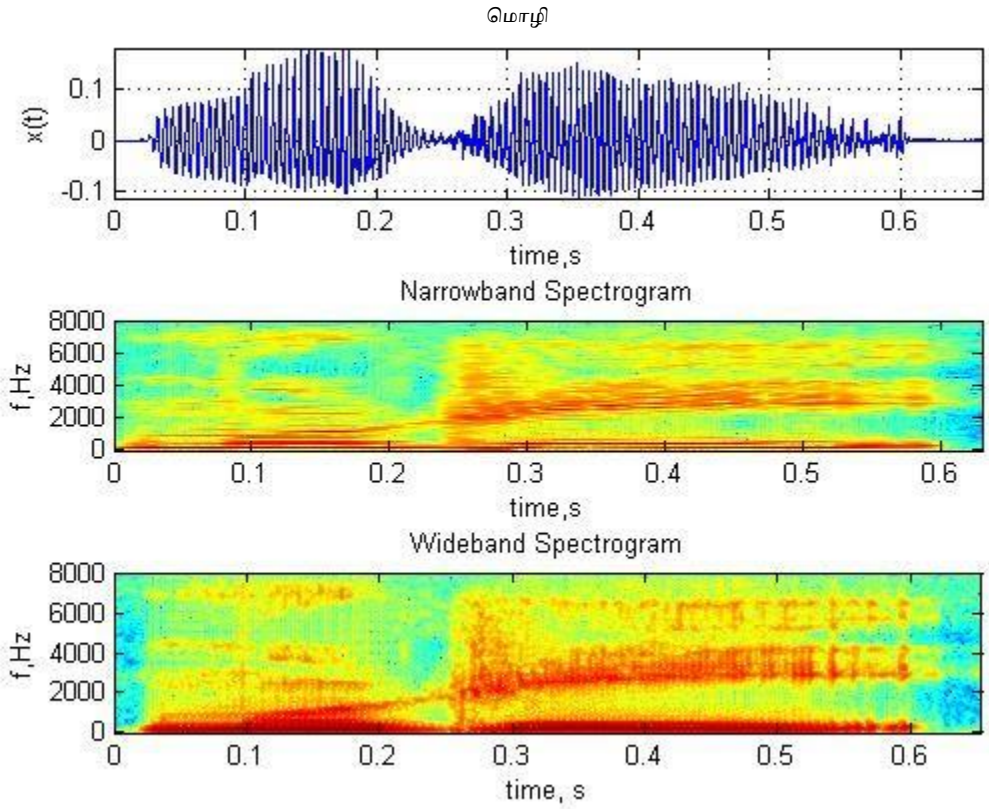
3.4.2.2.14. வ்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

வாலிபம்



3.4.2.2.15. ழ்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

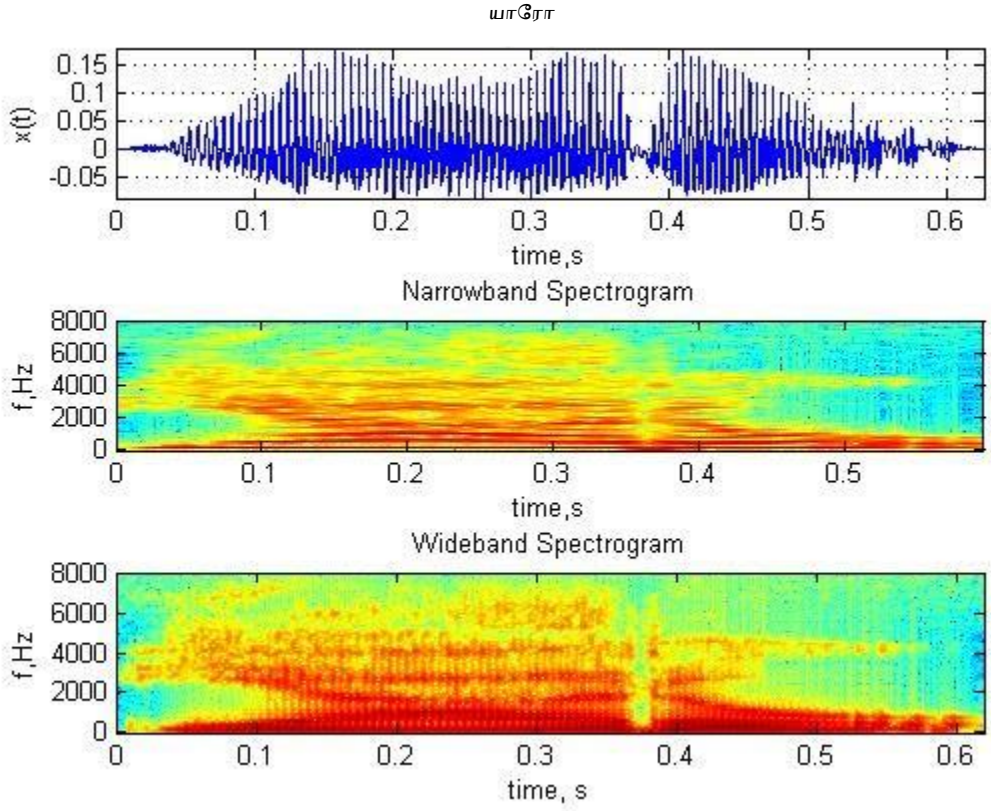
மொழி



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

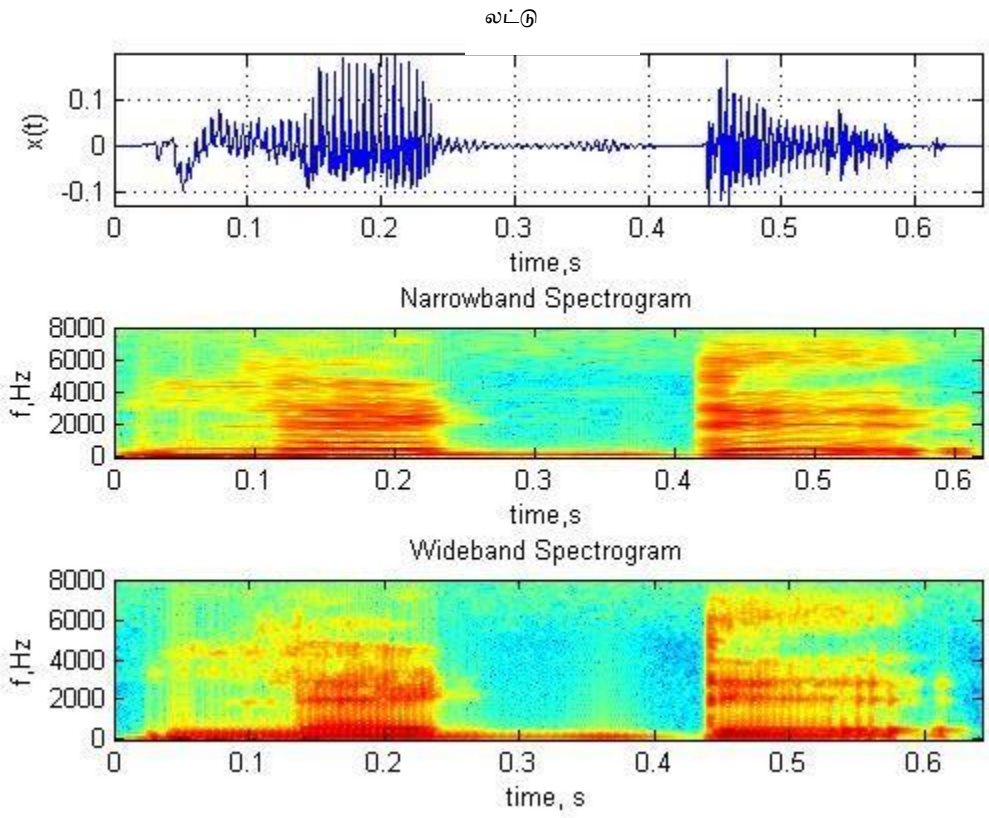
3.4.2.2.16. ய்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

யாரோ



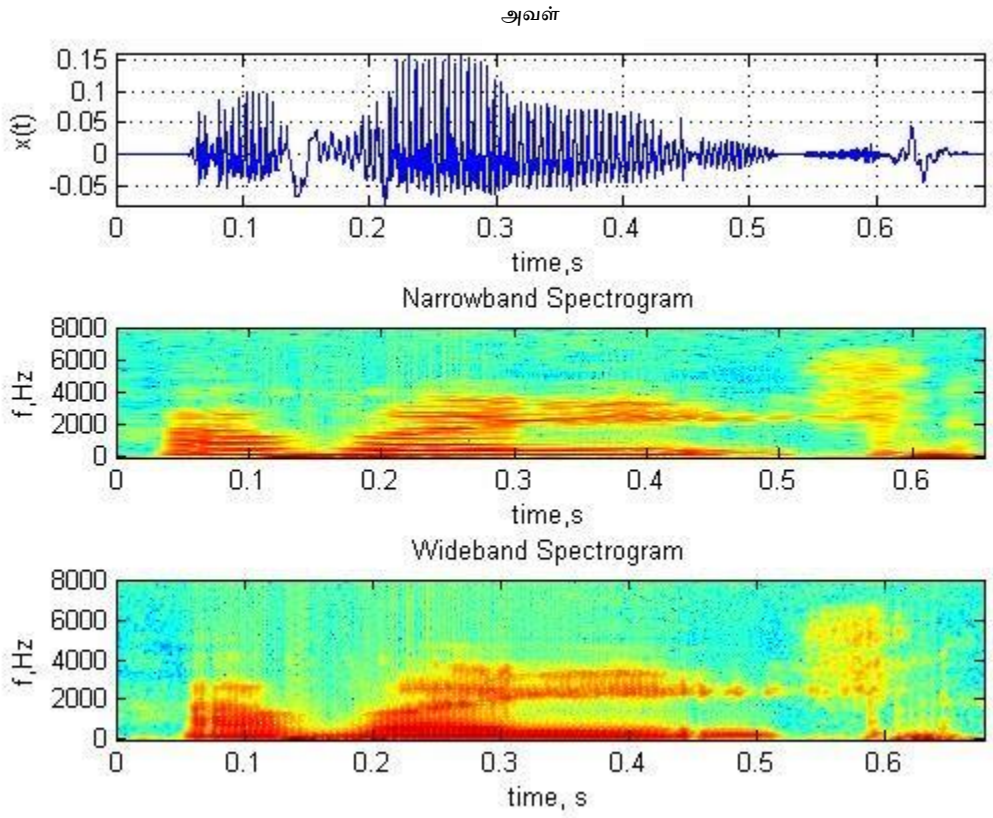
3.4.2.2.17. ல்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

லட்டு



3.4.2.2.18. ள்-இன் ஒலியியக்கப் பண்புக்கூறுகள் வெளிப்படுத்தும் நிறமலைவரைவு

அவள்



3.5. உரையிலிருந்து பேச்சாக்கம்

பேச்சு சொல் பேச்சின் சிறு அலகுகளால் ஒருங்கிணைக்கப்பட்டது என்ற ஒலி அடிப்படையிலான எழுதும் ஒழுங்குமுறையில் உள்ளுறையும் கருத்து ஒலியனியலின் எல்லாத் தற்காலக் கோட்பாடுகளையும் உள்ளடக்கும் யுஆர்-கோட்பாடு ஆகும். நாம் இங்கு தற்காலப் பேச்சறிதல் மற்றும் பேச்சுருவாக்கத் தொழில்நுட்பத்தையும் கணினி ஒலியனியல் (computational phonology) என்ற தொடர்புள்ள மொழியியலின் கிளைளையும் அறிந்துகொள்ள அடிப்படை உள்ளறிவையும் வழிமுறை வரைவுகளையும் அறிமுகப்படுத்த வேண்டும்.

தானியங்கு பேச்சறியும் முக்கியமான செயல்பாடு ஒலியியக்க அலைவடிவை உள்ளீடாக ஏற்றுச் சொற்களின் கோர்வையை வெளியீடாக உற்பத்தி செய்யும். எதிரிடையாக உரையிலிலிருந்து பேச்சுருவாக்கம் என்ற முக்கியச் செயல்பாடு உரைச் சொற்களின் கோர்வையை உள்ளீடாக ஏற்று ஒலியியக்க அலைகளை வெளியீடாகப் பெறுவதாகும். பேச்சைப் புரிந்து கொள்ளுதல் மற்றும் உருவாக்குதல் இவற்றின் பயன்பாடு பல்வகைப்பட்டதாகும்: தானியங்கு கேட்டெழுத்துதல் (automatic dictation), ஒலிபெயர்ப்பு, கணிப்பொறிக்குப் பேச்சு அடிப்படையிலான இடைமுகங்கள் மற்றும் தொலை பேசிகள், பேச்சுக் குறையுள்ளவர்களுக்குக் குரல்சார்ந்த உள்ளீடு மற்றும் வெளியீடு என்பனவாகும்.

இங்கு நாம் பேச்சறிவான் (speech recognizer) மற்றும் உரையிலிருந்து பேச்சு (text-to-speech) ஒழுங்குமுறைகள் என்பனவற்றில் கவனம் செலுத்துவோம்: எவ்வாறு சொற்கள் ஒலிகள் (phones) என்ற தனிப்பட்ட பேச்சு அலகின் அடிப்படையில் உச்சரிக்கப்படுகிறது? பேச்சறியும் ஒழுங்குமுறை அது புரிந்துகொள்ள இயலும் ஒவ்வொரு சொல்லுக்கும் உச்சரிப்பை வேண்டும். உரையிலிருந்து பேச்சு ஒழுங்குமுறை ஒவ்வொரு சொல்லுக்கும் ஒரு உச்சரிப்பை வேண்டும். நாம் இங்கு உச்சரிப்பை விளக்க ஒலி நெடுங்கணக்குகளை (phonetic alphabets) ஒலியியல் (phonetics) களத்தின் பகுதியாக அறிமுகப்படுத்த வேண்டி வரும். மேலும் நாம் ஒலிப்புமுறை ஒலியியல் (articulatory phonetics) என்பதையும் அறிய வேண்டி வரும். இது எவ்வாறு வாயிலுள்ள ஒலிப்பான்களால் பேச்சொலிகள் உருவாக்கப்படுகிறது என்பதை ஆயும்.

ஒரு குறிப்பிட்ட ஒலி எப்பொழுதும் ஒவ்வொரு சூழலிலும் ஒரேமாதிரி உச்சரிக்கப்படுகிறது என்றிருக்குமானால் உச்சரிப்பை மாதிரிப்படுத்துவது எளிதாகும். ஆனால் இது அவ்வாறில்லை. எடுத்துக்காட்டாக, தமிழில் [த்] என்ற ஒலி வேறுப்பட்ட ஒலிச் சூழல்களில் மிக வேறுபட்டு மெய்ப்படுத்தம் செய்யப்படுகிறது. மொழியியலில் ஒலியனியல் (phonology) களம் வேறுபட்ட சூழலில் ஒலிகள் வேறுபட்டு மெய்படுத்தம் செய்யப்படுகின்றது என்பதையும் எவ்வாறு ஒலிகளின் இவ்வொழுங்குமுறை மீதி இலக்கணத்துடன் தொடர்புடையது என்பதையும் விளக்குகிறது. நாம் இவ்வேறுப்பட்ட மெய்ப்படுத்தங்களை விளக்க ஒலியனியல் விதிகளை (phonological rules) எழுத வேண்டும்.

கணினி வழி ஒலியனியல் ஆய்வு கணினி ஒலியனியல் (computational phonology) என்று அறியப்படும். கணினி ஒலியனியலின் முக்கியமான பகுதி ஒலியனியல் விதிகளை மாதிரிப்படுத்தும் கணினி இயங்குமுறைகளின் ஆய்வாகும். நாம் எவ்வாறு எழுத்துக்கூட்டல் மாற்றிகள் (spelling-rule transducers) ஒலியனியலை மாதிரிப்படுத்த இயலும் என்று காட்ட வேண்டும். பின்னர் ஒலியனியல் கற்றலின் (phonological learning) கணினி மாதிரிகளை விளக்க வேண்டும்: எவ்வாறு ஒலியனியல் விதிகள் இயந்திரக் கற்றல் வழிமுறைவரைவுகளால் தானாக ஊக்கப்படுத்த இயலும். இறுதியாக நாம் உரையிலிருந்து பேச்சு ஒழுங்குமுறைகளில் ஒரு முக்கியமான சிக்கலுக்கு வேண்டி ஒலியனியலின் மாற்றி அடிப்படையிலான மாதிரியைப் (transducer-based model) பிரயோகிக்க வேண்டும்: எழுத்துக்களின் கோர்வைகளிலிருந்து ஒலிகளின் கோர்வைகளைப் பொருத்துதல். நாம் முதலில் பெரிய ஒலிப்பு அகராதியை உருவாக்குவதில் உள்ள சிக்கல்களை அளவிட வேண்டும். பின்னர் எவ்வாறு மாற்றி அடிப்படையிலான சொற்களஞ்சியங்களும் எழுத்துக்கூட்டல் அலகுகளும் எழுத்திலிருந்து உச்சரிப்புக்குப் பொருத்துவதற்கு உச்சரிப்புகளால் பெரிதுபடுத்தப்பட வேண்டும் என்று காட்ட வேண்டும்.

3.5.1. உரையிலிருந்து பேச்சு

உரையிலிருந்து பேச்சு என்ற கலைச் சொல் உள்ளீட்டு உரையைப் பேச்சு கூற்றாக மாற்றுவதைக் குறிக்கிறது. உள்ளீட்டு உரையில் சொற்கள், வாக்கியங்கள், பத்திகள், எண்கள் மற்றும் சுருக்கங்கள் கூட உட்படுத்தப்படும். உரையிலிருந்து பேச்சு மாற்று செயல்முறை உரையை அடையாளம் காண வேண்டும். எந்த தெளிவற்ற தன்மையும் இல்லாமல், அதனுடன் தொடர்புடைய ஒலி வெளியீட்டை ஏற்றுக்கொள்ளத்தக்க வகையில் உருவாக்கவேண்டும். இதன் பொருள் உரையிலிருந்து பேச்சு இயந்திரத்தின் வெளியீட்டின் தரம் இயற்கையான பேச்சுக்கு முடிந்தவரை நெருக்கமாக இருக்கும்படி செய்யப்பட வேண்டும்.

3.5.2. ஒலியனியல்

பேச்சொலிகளின் அலகுகளை அசைகளாகவும் பிற பெரிய அலகுகளாகவும் ஒழுங்குபடுத்துவதை ஆய்வது ஒலியனியலாகும். ஒரு மொழியின் ஒலியனியல் அம்மொழியில் நிகழும் ஒலிகளின் ஒழுங்குமுறைகள் மற்றும் அமைப்பொழுங்குகள் இவற்றின் வர்ணனையாகும். எவ்வாறு ஒலிகள் எல்லாம் ஒருங்கிணைந்து ஒலியன்களாகக் கருதப்படுகின்றன, ஒலியன்கள் எவ்வாறான வரிசைக் கிரமத்தில் வந்து உருபன்களாகவோ சொற்களாகவோ உருவாக்கம் பெறுகின்றன என்பதை ஒலியனியல் விவரிக்கின்றது. ஒரு மொழியின் ஒலி ஒழுங்குமுறையை ஆயும் மொழியியலின் ஒரு கிளைப்பாடம் ஒலியனியலாகும். அதாவது ஒரு மொழியின் ஒலியன்களைக் கண்டுபிடித்து அவற்றின் வரன்முறை, அமைப்பொழுங்கு, செயற்பாடு போன்ற செய்திகளை விளக்குவது ஒலியனியலாகும் (phonology). ஒரு மொழியின் கூற்றுகளை வேறுபடுத்தும் குறைந்த செயற்பாட்டு அலகு ஒலியனாகும் (phonemes). மனித பேச்சு உறுப்புகள் உற்பத்தி செய்ய இயலும் ஒலிகளின் விரிந்த பரப்பெல்லையில் ஒத்தறி அடிப்படையில் குறைந்த எண்ணிக்கையிலான ஒலிகள் தான் எந்த ஒரு மொழியிலும் பயன்படுதப்படும். ஒலிகள் முரண்படும் ஒழுங்கு முறைகளாக அமைக்கப்படுகையில் அவை ஒலியன்கள் எனப்படும். ஒரு மொழியின் ஒலியன்களை நிறுவி அவற்றின் மாற்றொலிகளைக் கண்டுபிடித்து வருணனை விளக்கம் தருவதும் ஒலியன்களின் வரன்முறையை விளக்குவதும் ஒலியனியலியலின் முக்கிய நோக்கங்களாகும். ஒலி, ஒலியன், மாற்றொலி இவற்றிற்கிடையில் வேறுபபாடுகள் உள்ளன. ஒலியன்களைக் கண்டுபிடிப்பதற்கென்று வழிமுறைகள் இருக்கின்றன. ஒரு மொழியின் ஒலியன்கள் ஒரு குறிப்பிட்ட வரன் முறை அமைப்பைக் கொண்டிருக்கும். ஒலியன்களை விளக்க பல கோட்பாட்டுகள் உள்ளன. ஒலி (phone), ஒலியன் (phoneme), மாற்றொலி (allophone) இவற்றின் வேறுபாடுகள் பற்றி அறித்திருத்தல் அவசியம் ஆகும். எந்த ஒலியும் (அது மொழிசார்ந்ததாய் இருந்தாலும் இல்லாவிட்டாலும்) ஒலிதான். அதை ஒலியன் என்று குறிப்பிட இயலாது. ஒலியன்

என்பது ஒரு மொழி சார்ந்த ஒலி. இந்த ஒலி கூற்றுக்களை வேறுபடுத்தும் செயல்பாட்டை மொழியில் செய்யும். எடுத்துக்காட்டாக *பல்* என்பது *கல்* என்பதிலிருந்து ப்-க் என்ற ஒலி வேறுபாட்டால் வேறுபடும் சொற்களாகும். எனவே *ப்* என்பதும் *க்* என்பதும் ஒலியன்கள் ஆகும். ஒரு மொழிக்கு எழுத்து வடிவமும் ஒலியன்களால் தான் தீர்மானிக்கப்படும். ஒலியன்களைப் பற்றி படிப்பது தான் ஒலியனியல். பேச்சொலிகளை அவை கூற்றுக்களை வேறுபடுத்தும் செயல்பாட்டின் அடிப்படையில் ஒலியன்களாக இனங்கண்டு அவற்றின் வருகையிடங்கள், சேர்க்கைகள், அதனால் அமையும் அசைகளின் ஒழுங்குமுறைகள் இவற்றை ஆய்ந்து விளக்குவதே ஒலியனியல் எனப்படும். எடுத்துக்காட்டாக பின்வரும் சொற்களை அவற்றின் உச்சரிப்பின் அடிப்படையில் எழுதுவது பின்வருமாறு அமையும்.

1. பல் [pal] - [p]
2. கல் [kal] - [k]
3. பத்து [pattu] - [p]
4. தம்பி [tambi] - [b]

முதல் இரண்டு எடுத்துக்காட்டுகளில் [p], [k] என்பன முரண்பட்டு கூற்றுகளை அதாவது இரண்டு பொருள்கொண்ட சொற்களை வேறுபடுத்துகின்றன. எனவே இவை இரண்டும் ஒலியன்களாகும். 3-ஆவது மற்றும் 4-ஆவது எடுத்துக்காட்டுகளில் [p], [b] என்பன முரண்பட்டு வருவதில்லை. [p] என்பது சொல்லின் முதலிலும் [b] என்பது [அ] என்ற ஒலியை அடுத்தும் வருகின்றன. [m] என்ற மூக்கொலி குரலொலியாகும். இவ்வொலியானது தன்னை அடுத்து வரும் [p] என்ற குரலிலா ஒலியை குரலொலியாக அதாவது [b] ஆக மாற்றியது என நாம் விளங்கிக் கொள்ளலாம். அதாவது [b] என்பதும் [p] என்பதும் /p/ என்ற ஒலியனின் மாற்றொலிகள் எனலாம். [b], [p] என்பன இரு வேறுபட்ட (முரண்படாத) இடங்களில் வருகின்றன. இவ்வாறு இரு ஒலிகள் வேறுபட்ட இடங்களில் வருவது துணைநிலை வருகை முறை எனப்படும். [p], [k] என்பன ஒன்றையொன்று இடம்பெயர்ப்பதுபோல் ஒரே இடத்தில் வருவது வேற்றுநிலை வருகை முறை (contrastive distribution) எனப்படும். மாறாக ஒரு ஒலியனின் மாற்றொலிகள் ஒன்றின் சூழலில் மற்றொன்று வராமல் வேறுபட்ட சூழல்களில் வருதல் துணைநிலை வருகை முறை எனப்படும். எடுத்துக்காட்டாக, தமிழில் *த்* என்ற ஒலி சொல் முதலிலும் தகரத்திற்கு முன்னரும் மட்டுமே வரும். இதன் இணையான *ன்* என்ற ஒலி இவ்விரு சூழல்களைத் தவிர்த்து பிற சூழல்களில் வரும். என்வே *த், ன்* என்பவைகளை ஒரு ஒலியனின் மாற்றொலிகளாகக் கொள்ள இயலும். இருப்பினும் வேறு காரணங்களால் அவை இரண்டு எழுத்துக்களால் தமிழில் மெய்ப்படுத்தம் செய்யப்பட்டுள்ளன.

ஒலியன் என்பது ஒலிகளை எழுத்து வடிவத்திற்கு கொண்டு வருவதற்கு தேவைப்படும் அலகு ஆகும். ஒரு மொழியில் ஒலியன்களைத் தவிர இசைமை (pitch), அழுத்தம் (stess), இசையோட்டம் (intonation), சுரம் (tone) போன்ற ஒலிக் கூறுகளும் முரண்பாடுகளை உருவாக்கலாம். இவை மீக் கூற்று ஒலியன்கள் (supresegmental phonemes) எனப்படும்.

ஒலியன்களைக் கண்டுபிடிக்கும் முறை (method of finding phonemse) ஒலியன்களைப் பற்றிய ஆய்வுக்கு வரிவடிவாக்கம் அல்லது எழுத்துப் பெயர்ப்பு செய்யபட்ட மொழித்தரவு தேவை. எழுதப்படாத பழங்குடி மக்களின் மொழிகளை அல்லது கிளை மொழிகளை ஆயும் மொழியியலார் தாங்கள் பெற்ற ஒலியியல் பயிற்சியின் பயனாய்த் தகவலாளி தரும் கூற்றுகளை எழுத்துப் பெயர்பு (transcription செய்வர்). எழுத்துப்பெயர்ப்பு செய்பவரின் குறைபாடு காரணமாக தரவின் பதிவில் பிழைகள் ஏற்படலாம். இத்தகைய பிழைகள் ஒலியியல் பயிற்சியின் குறைபாடு காரணமாகவோ செறிவு எழுத்துப்பெயர்ப்பு (narrow transcription) காரணமாகவோ செறிவல்லாத எழுத்துப்பெயர்ப்பு (broad transcription) காரணமாகவோ ஒலிகளைத் தவறாகப் பிரிப்பதனாலோ ஏற்படலாம். ஒலியனியல் ஆய்வுக்கு முன்னர் இக்குறைபாடுகள் மொழித்தரவிலிருந்து நீக்கப்படவேண்டும். அதன் பின்னர் ஒலியனியல் ஆய்வு அதாவது ஒலியன்களைக் கண்டுபிடிக்கும் செயல்முறைகள் கடைபிடிக்கப்படும்.

3.5.3. ஒலியன்களும் ஒலியன் விதிகளும்

உரையிலிருந்து பேச்சாக்கத்தில் எழுத்துக்களால் அல்லது எழுத்துக்களின் கோவைகளால் ஆன உரையானது ஒலிவடிவில் அல்லது ஒலிகளின் கோவையாக மாற்றப்படவேண்டும். முதலில் எழுத்துக்கள் ஒலியன்களாக (phonemes) மாற்றப்படவேண்டும். தமிழ் போன்ற மொழிகளில் இது எளிது ஏனென்றால் தமிழ் எழுத்துக்கள் தமிழ் ஒலியன்களின் நேரடி வடிவம் தான். ஆனால் ஆங்கிலம் போன்ற மொழிகளில் எழுத்துவடிவமும் ஒலியன் வடிவமும் வேறாகும். எழுத்துக்களின் கோவைகளுக்கும் ஒலியன்களின் கோவைகளுக்கும் உள்ள தொடர்பு நேரடியானதல்ல. எனவே ஆங்கிலம் போன்ற மொழிகளில் எழுத்து வடிவம் அல்லது எழுத்துகளின் கோவை ஒலியன்களின் கோவையாக மாற்றப்படவேண்டும். எனவே எழுத்துகளிலிருந்து ஒலியன்களாக மாற்றும் விதிகள் ஆங்கிலத்தில் தேவை. தமிழில் இது எளிதாகும்.

ஒலியன்களின் கோவைகளாக மாற்றப்பட்ட உரை ஒலிகளின் கோவைகளாக மாற்றப்படவேண்டும். ஒலியன்கள் வேறுபட்ட சூழல்களில் வேறுபட்டு உருவாக்கப்படும். ஒலியன்கள் தமக்குள் சூழல் அடிப்படையில் மாற்றொலிகளின் (allophones) ஒரு குழுமத்தால் உருப்படுத்தம் செய்யப்படும். ஒலியனுக்கும் அதன் மாற்றொலியன்களுக்கும் இடையில் உள்ள

உறவு ஒலியன்களின் விதியால் பெறப்படும். எடுத்துக்காட்டாக, தமிழில் /க்/ என்ற ஒலியனை உருப்படுத்தம் செய்யப்படும் மூன்று மாற்றொலியன்களைப் பின்வருமாறு ஒலியனியல் விதியால் விளக்கலாம்:

k ⟶ g/N

k ⟶ x/V-V

k ⟶ k/ பிற சூழல்களில்

அதாவது மூக்கொலி முன் கடையண்ணக் குரலிலா ஒலியானது கடையண்ணக் குரல் ஒலியாகவும் இரு உயிர்களுக்கிடையில் கடையண்ண உரசொலியாகவும் பிற சுழல்களில் குரலிலா அடைப்பொலியாகவும் உருப்படுத்தம் பெறும். தமிழில் /ப்/ என்ற ஒலியனுக்கு நான்கு மாற்றொலிகள் உள்ளன. இம்மாற்றொலிகள் உச்சரிப்பால் வேறுபடும் மற்றும் இவை தமது வருகையால்/வினியோகத்தால் அல்லது கூற்றில் அவை வரும் இடத்தால் வேறுபடும். இதுபோன்று /த்/, /ட்/, /ச்/, /க்/ என்ற ஒலியன்களுக்கு முறையே நான்கு மாற்றொலிகள் உள்ளன. /ற்/ என்பதற்கு மூன்று மாற்றொலிகள் உள்ளன. பிற மெய்யொலிகள் ஒரேயொரு மாற்றொலியைக் கொண்டுள்ளன. மெய்யொலியன்களைப் போலவே உயிரொலியன்களும் மாற்றொலிகளைக் கொண்டுள்ளன: /இ/ என்பது மூன்று மாற்றொலிகளை கொண்டுள்ளது; /ஈ/ ஒரேயொரு மாற்றொலி கொண்டது; /எ/ இரண்டு மாற்றொலிகளைக் கொண்டது; /ஏ/ ஒரேயொரு மாற்றொலியைக் கொண்டது; /அ/ இரண்டு மாற்றொலிகளைக் கொண்டது; /ஆ/ ஒரேயொரு மாற்றொலியைக் கொண்டது; /உ/ இரண்டு மாற்றொலிகளைக் கொண்டது; /ஊ/ ஒரெயொரு மாற்றொலியைக் கொண்டது; /ஒ/ இரண்டு மாற்றொலியைக் கொண்டது. /ஓ/ ஒரேயொரு மாற்றொலியைக் கொண்டது.

உரையிலிருந்து பேச்சாக்கத்தில் ஒலியன்களின் மாற்றொலிகள் கவனத்தில் கொள்ளப்படவேண்டும். அவ்வாறு செய்யும் போது தான் பேச்சாக்கம் இயற்கையானதாக இருக்கும்.

3.5.4. ஒலியனியல் விதிகளும் முற்றுநிலை மாற்றிகளும்

ஒலியனியல் விதிகளை முற்றுநிலை மாற்றிகளால் நடைமுறை படுத்தலாம். ஒலியனியல் விதிகளை மெய்ப்படுத்தம் செய்ய முற்றுநிலைத் தானியங்கியைப் பலவழிகளில் பயன்படுத்தும் கணினி ஒலியனியலின் வேறுப்பட்ட மாதிரிகள் இருக்கின்றன. இவற்றில் ஒன்று நாம் கணினி உருபனியலில் பார்க்க இருக்கும் கொஸ்கென்னியின் இருமட்ட உருபனியலாகும் (two-level morphology of Koskenniemi (1983)). இருமட்ட உருபனியல் சொல் வடிவுக்கும் புற வடிவுக்கும் இடையே உள்ள பொருத்தத்தின் மீது முற்றுநிலை நல்லுருவாக்கக் கட்டுப்பாடுகளாக மாதிரி செய்யும் உருபனியல்/ஒலியனியல் கோட்பாடாகும். கணினி ஒலியனியலின் உயர்வான சிக்கல்களை நேரிட ஒலியனியலின் இருமட்ட அல்லது முற்றுநிலை மாதிரியின் கூடுதல் சிறப்பான உபாயம் பயன்படுத்தப்படும். ஒலியனியல் விதிகளைப் பொறியால் கற்கச் செய்யலாம். ஒரு பரப்பிலிருந்து சிறிது தரவும் பிற தகவல்களும் தரப்படுகையில் அது அப்பரப்பிற்குத் தானாக ஒரு மாதிரியைத் தூண்டுவது பொறி கற்றல் ஒழுங்குமுறையின் வேலையாகும். இவ்வாறு ஒலியனியல் விதிகளைக் கற்கவேண்டி ஒரு ஒழுங்குமுறைக்கு அதிலிருந்து தூண்ட சொற்களின் (புற

வடிவங்களின்) குறைந்தது ஒரு குழும விதிகள் தரப்படும். ஒரு கண்காணிக்கப்பட்ட வழிமுறைவரைவு (supervised algorithm) இத்தரவின் சிறுபகுதிக்கு விடைகள் தரப்படுகையில் இவ்விடைகளைப் பயன்படுத்தி அது முன்னர் பார்க்காத புதிய தரவைப் பொதுமையாக்கம் செய்யவியலும் ஒரு மாதிரியைத் தூண்டும். கண்காணிக்கப்படாத வழிமுறைவரைவு (unsupervised algorithm) இதைத் தரவிலிருந்தே செய்கின்றது. கண்காணிக்கப்படாத வழியமைப்புகள் வகுப்புகளுக்குச் சரியான புலக்குறிகளைப் பார்க்காமல் இருந்தாலும், அவைகள் உருவாக்க வேண்டிய விதிகளின் அல்லது மாதிரிகளின் இயல்பைப் பற்றி குறிப்பைத் தர இயலும். எடுத்துக்காட்டாக, தானியங்கி வடிவில் மாதிரிகள் இருக்கும் என்ற அறிவு ஒருவகைக் குறிப்பாகும். இம்மாதிரியான குறிப்புகள் கற்றல்சார் தவறானகருத்துகள் (learning bias) எனப்படும்.

3.5.5. உரையிலிருந்து பேச்சுக்குவேண்டி உரையை ஒலியன்களுக்குப் பொருத்துதல்

நாம் ஒரு மொழியின் ஒலிகளின் பட்டியலைத் தெரிந்துக் கொள்ளவும் எவ்வாறு ஒலியனியல் விதிகளை மாதிரிப் படுத்துவது என்பதை அறிந்துகொள்ளவும் செய்தால் எழுத்திலிருந்து அல்லது உரைச் சொல்லிருந்து அதன் உச்சரிப்புக்குப் பொருத்தும் சிக்கலை ஆயத் தயார் என்பதாகும்.

3.5.5.1. உச்சரிப்பு அகராதிகள்

உச்சரிப்பு அகராதிகள் உரையிலிருந்து பேச்சுக்குப் பொருத்த மிகப் பயனுள்ள கருவியாகும். எளிய உச்சரிப்பு அகராதிகள் சொற்களின் பட்டியலையும் அவற்றின் உச்சரிப்பையும் தரும் (Jurafsky and Martin, 2000: 120).

Word	Pronunciation	Word	Pronunciation
cat	[kæt]	goose	[gus]
cats	[kæts]	geese	[gis]
pig	[pɪg]	hedgehog	[‘hɛʤ.hɔg]
pigs	[pɪgz]	hedgehogs	[‘hɛʤ.hɔgz]
fox	[fax]		
foxes	[‘fak.sɪz]		

PRONLEX, CMUdict, மற்றும் CELEX என்ற இணையத்தள உச்சரிப்பு அகராதிகள் பொதுவாகப் பயன்படுத்தப்படுகின்றன. இவை பேச்சைத் தெரிந்து கொள்கைக்குப் பயன்படுத்தப்படுகிறது. இவற்றைப் பேச்சு கூட்டிணைப்புக்கும் பயன்படுத்த வேண்டி சுவீகரித்துக் கொள்ளலாம்.

3.5.5.2. அகராதியை நோக்குவதற்கு அப்பாற்பட்ட உரை ஆய்வு

ஆரம்பகால உரைகளிலிருந்து பேச்சாய்வுகள் உரையிலிருந்து ஒலிகளுக்குப் பொருத்த உச்சரிப்பு அகராதிகளைச் சார்ந்திருந்தன. முன்னர் கூறியபடி உரைகளிலிருந்து ஒலிகளாகப் பொருத்தமாற்றம் செய்யும் ஒரு வழி ஒவ்வொரு சொற்களின் உச்சரிப்பையும் உச்சரிப்பு அகராதியைப் பார்த்தறிந்து ஒலிகளின் கோர்வையை அகராதியிலிருந்து படிப்பதாகும். இந்த நெறிமுறை ஏற்கெனவே அகராதியில் தரப்பட்ட சொற்களுக்குப் பொருத்தும். ஆனால் பலச்சொற்கள் முன்னரே அகராதியில் தரப்படாததற்கான சூழல் சாத்தியம் இருக்கின்றன. இத்தகைய நிலையில் நாம் அகராதியில் தரப்படாத சொற்களின் உச்சரிப்பைத் தெரிந்துக்கொள்ள நிகழ்வுசாத்திய நுட்பங்களைப் (probabilistic techniques) பயன்படுத்த வேண்டிவரும்.

3.5.5.3 முற்றுநிலைமாற்றி அடிப்படையில் உச்சரிப்பு அகராதி

உரையிலிருந்து பேச்சு ஒழுங்குமுறைகளின் தொடக்கக்கால உச்சரிப்பு மாதிரி எழுத்திலிருந்து ஒவ்வொரு விதியும் எவ்வாறு ஒரு எழுத்து அல்லது எழுத்துக்களின் இணைப்பு ஒலிகளாக பொருத்தமாற்றம் செய்யப்பட்டன என்று குறிப்பிட்டுக் காட்டியது. கீழே விற்றன் (Witten, 1982) என்பவரின் விதி அடிப்படை வடிவு தரப்பட்டுள்ளது (Jurafsky and Martin, 2000: 125):

Fragment	Pronunciation
- p -	[p]
- ph -	[f]
- phe -	[fi]
- phes -	[fiz]
- place -	[pleɪs]
-placi -	[pleɪsi]
-plement -	[plɪmɛnt]

ஸ்ஃப்ரோட் (Sproat) மற்றும் அவருடன் சேர்ந்த பிறரும் உரையிலிருந்து பேச்சுக்கு மாற்றிகளைப் பயன்படுத்தினார்கள். நாம் அவர்களின் மாற்றி அணுகுமுறையை ஐந்து பிரிவுகளாகப் பிரிக்கலாம்:

1. அகராதியிலுள்ள தனிச் சொற்களுக்கும் உருபன்களுக்கும் உச்சரிப்பை உருப்படுத்தம் செய்யும் முற்றுநிலை மாற்றி
2. உருபன்களின் சாத்தியமானத் தொடர்ச்சியை உருப்படுத்தம் செய்யும் முற்றுநிலைத் தானியங்கிகள்
3. ஒவ்வொரு உச்சரிப்பு விதிகளுக்கும் தனியான முற்றுநிலை மாற்றி
4. பெயர்களின் மற்றும் சுருக்கப்பெயர்களின் உச்சரிப்பை மாதிரிப்படுத்தப் பயன்படுத்தப்படும் உள்ளுணர்வு மற்றும் எழுத்திலிருந்து-ஒலி விதிகள்/ மாற்றிகள்

5. எந்த பிற அறியாத சொற்களுக்கும் வழுநிலை எழுத்திலிருந்து-ஒலி விதிகள்/மாற்றிகள் உரையிலிருந்துப் பேச்சுக்கான உரையிலிருந்து ஒலிக்குப் பொருத்தமாற்றம் (Jurafsky and Martin, 2000:127):

Orthographic Lexicon	Lexicon
Regular Nouns	
cat	c/k a|æ t|t
fox	f|f o|ɑ x/ks
dog	d|d o|ɑ g|g
Irregular Singular Nouns	
goose	g|g oo|u s|s e|ɛ
Irregular Plural Nouns	
go:e o:e se	g| g oo|u: ee|i s|s e|ɛ

மேற்சொன்னப் படத்தில் அகராதிக்கு இரண்டு நிலைகள் இருக்கின்றன. ஒன்று உள்ளுறையும் சொல்சார் மட்டம்; மற்றொன்று இடைப்பட்ட மட்டம். இதில் இடைப்பட்ட மட்டத்தைப் புற மட்டத்திற்கு பொருத்தமாற்றம் செய்ய எழுத்துக்கூட்டல் விதிகளையும் உச்சரிப்பு விதிகளையும் பயன்படுத்தும் மாற்றிகளைச் சேர்க்க வேண்டும். இது வேறுபட்ட எழுத்துகூட்டல் விதிகளையும் ஒலியன் விதிகளையும் உள்ளடக்கும். சொற்களஞ்சியம் மற்றும் ஒலியனியல் விதிகள் மற்றும் வரிவடிவ விதிகள் என்பனவற்றைச் சொல் உருப்படுத்தத்திற்கும் புற உருப்படுத்தத்திற்கும் இடையில் பொருத்த மாற்றம் செய்யப் பயன்படுத்தலாம். சொற்களஞ்சிய முற்று நிலை மாற்றிகள், அடிகள் மற்றும் ஒலியனியல் பண்புக் கூறுகளைக் கொண்ட சொல் மட்டத்திற்கும் உருபன்களின் எளிய இணைப்பை உருப்படுத்தம் செய்யும் இடைப்பட்ட மட்டத்திற்கும் இடையில் பொருத்தமாற்றம் செய்கிறது. பின்னர் ஒரு தனி எழுத்து விதிக் கட்டுப்பாடு அல்லது தனி ஒலியனியல் கட்டுப்பாடு இவற்றில் ஒன்றை உருப்படுத்தம் செய்யும் ஒருகுழும முற்றுநிலை மாற்றிகள் இடை மட்டம் மற்றும் புறமட்டம் இவற்றிற்கிடையில் பொருத்தமாற்றம் செய்ய இணையாகச் செயல்படுகின்றது. ஒவ்வொரு மட்டத்திற்கும் வரிவடிவ மற்றும் ஒலியனியல் உருப்படுத்தங்கள் இருக்கின்றன. உள்ளீடு சொல் வடிவில் இருக்கும் உரையிலிருந்து-பேச்சு பயன்பாடுகளுக்கு (எ.கா. ஒழுங்குமுறை சொல்லின் சொல்சார் அடையாளம், அதன் சொல்வகைப்பாடு, அதன் திரிபு இவற்றை அறியும் உரை ஆக்கத்திற்கு) முற்றுநிலை மாற்றிகள் சொல் வடிவிலிருந்து புற உச்சரிப்புக்குப் பொருத்த மாற்றம் செய்ய இயலும். உள்ளீடு புற எழுத்துக்கூட்டலாக இருக்கும் உரையிலிருந்து-பேச்சுப் பயன்பாடுகளுக்கு (எ.கா. “உரையை

உரக்கப் படிக்கும்" பயன்பாடுகளுக்கு) முற்றுநிலை மாற்றிகள் உள்ளுறையும் சொல் வடிவு வழி புற வரிவடிவத்திலிருந்து புற உச்சரிப்பை பொருத்தமாற்றம் செய்யவியலும்.

இறுதிச் சிக்கல் தலைப்பெழுத்தன்களைப் (acronyms) பற்றியதாகும். தலைப்பெழுத்தன்களைக் முற்றுப்புள்ளிகளுடன் (எ.கா. I.R.S.) அல்லது முற்றுப்புள்ளிகள் இல்லாமல் (எ.கா. IRS) எழுத்துக்கூட்டலாம். முற்றுப்புள்ளிகளுடன் உள்ள தலைப்பெழுத்தன்களை எழுத்துக்கூட்டி உச்சரிக்கலாம் (எ.கா [aɪɑrɛs]). பொதுவாக முற்றுப்புள்ளிகள் இல்லாமல் தோன்றும் தலைப்பெழுத்தன்களைச் (எ.கா. *AIDS*, *ANSI*, *ASCAP*) சொல்லாக எழுத்துக்கூட்டவோ உச்சரிக்கவோ செய்யலாம். எனவே *AIDS* எனதை *aid* என்ற வினையின் படர்க்கை இட வடிவு போன்றே உச்சரிக்கலாம். லிபர்மான் மற்றும் சர்ச் (Liberman and Church, 1992) சொற்களாக உச்சரிக்கப்படும் தலைப்பெழுத்தன்களின் ஒரு சிறிய அகராதியை வைத்திருக்கவும் மீதமுள்ளவற்றை எழுத்துக்கூட்டவும் யோசனை கூறுகின்றார்.

3.5.6. உரையிலிருந்து-பேச்சில் மீக்கூறு

முன்னர் கூறிய வரிவடிவிலிருந்து ஒலிக்குச் செல்லும் மற்றிச் செயல்பாங்கு பேச்சை உருவாக்கும் உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் பகுதிக்கு உள்ளீட்டின் முக்கியக் கூறை உருவாக்கும். உள்ளீட்டின் மற்றொரு முக்கிய பாகம் மீக்கூறின் குறிப்பீடாகும். மீக்கூறு என்ற கலைச்சொல் பொதுவாகச் சொற்களஞ்சியத்திலிருந்து ஆக்கப்பட்ட ஒலிகளின் தொடர்ச்சியால் விளக்கப்படாத ஒரு வாக்கியத்தின் உச்சரிப்பின் நோக்குகளைக் குறிப்பிடப் பயன்படும். மீக்கூறு ஒலிகளைக் காட்டிலும் நீண்ட மொழி அலகுகளின் மீது செயல்படும். எனவே இது மீக்கூறு நடப்பின் ஆய்வு என்று சிலநேரங்களில் அழைக்கப்படும். மீக்கூறுக்கு மூன்று முக்கிய ஒலியனியல் நோக்குகள் இருக்கின்றன: மேன்மை (prominence), அமைப்பு (structure) மற்றும் சுரம் (tone). மேன்மை அசையழுத்தத்தையும் (stress) ஒலியழுத்தத்தையும் (accent) உள்ளடக்கும். மேன்மை என்பது அசைகளின் தன்மையாகும். இது ஒரு அசை மற்றொரு அசையைக்காட்டிலும் மேன்மையுடையது என்று ஒத்தறி அடிப்படையில் விளக்கப்படும். உச்சரிப்புச் சொற்களஞ்சியங்கள் சொல்சார் அசையழுத்தத்தைக் குறியீடுசெய்யும். எடுத்துக்காட்டாக *table* என்பது அசையழுத்தத்தை அதன் முதல் அசையிலும் *machine* என்பது அதன் அசையழுத்தத்தை இரண்டாவது அசையிலும் கொண்டிருக்கும். *there*, *the* அல்லது *a* போன்ற செயல்பாட்டுச் சொற்கள் பொதுவாக ஒலியழுத்ததைக் கொண்டிருக்காது. சொற்கள் ஒன்றாக இணைக்கப்படுகையில், அவற்றின் ஒலியழுத்த அமைப்பொழுங்கு இணைந்து முழு வாக்கியத்திற்கும் ஒரு பெரிய ஒலியழுத்த அமைப்பொழுங்கை உருவாக்கும். எவ்வாறு ஒலியழுத்தங்கள் இணைகின்றன என்பதில் சீரானதன்மைகள் இருக்கின்றன. ஒலியழுத்தத்தின் இடம் கருத்தாடல் காரணிகளால் மிக வலுவாகப் பாதிக்கப்படும்.

சில சொற்கள் இயல்பாகவே ஒன்றாகக் குழுமுவதாகத் தெரிகின்றது என்ற அளவில் வாக்கியங்களுக்கு மீக்கூறு அமைப்பு இருக்கின்றது. சில சொற்களுக்கு அவற்றிற்கிடையில்

காணத்தக்க இடைவெளி அல்லது பிரிநிலை இருக்கின்றது. பெரும்பாலும் மீக்கூறு அமைப்பு மீக்கூறு தொடராக்கம் (prosodic phrasing) அடிப்படையில் விளக்கப்படும். இப்பெரிய மீக்கூறு அலகுகளுக்குப் பயன்படுத்தப்படும் கலைச்சொற்கள் இசையோட்டத் தொடர் (intonational phrase or IP) இசையோட்ட அலகு (intonational unit) மற்றும் சுர அலகு (tone unit) என்பனவற்றை உள்ளடக்கும்.

3.5.6.1. மீக்கூறின் ஒலியல் அல்லது ஒலியியக்கவியல் நோக்குகள்

மூன்று ஒலியனியல் காரணிகள் ஊடாட்டம் செய்யும் மற்றும் அவை பல வேறுபட்ட ஒலியியல் மற்றும் ஒலியக்கவியல் நடப்புகளால் மெய்ப்படுத்தம் செய்யப்படும். மேன்மையான அசைகள் மேன்மையற்ற அசைகளைவிட பொதுவாக ஓசையுடையதாகவும் நீண்டதாகவும் இருக்கும். மீக்கூறுத் தொடர் எல்லைகள் பெரும்பாலும் விட்டிசையாலும் எல்லைக்கு முன்னர் அசையின் நீட்சியாலும் சில சந்தர்ப்பங்களில் எல்லையில் இசைமையின் குறைப்பாலும் தொடாப்படும். இசையோட்டச் சுரம் அடிப்படை நிகழ்வெண் (F0) வெளிவரம்பில் வெளிப்படுத்தப்படுகின்றது.

3.5.6.2. பேச்சு உருவாக்கத்தில் மீக்கூறு

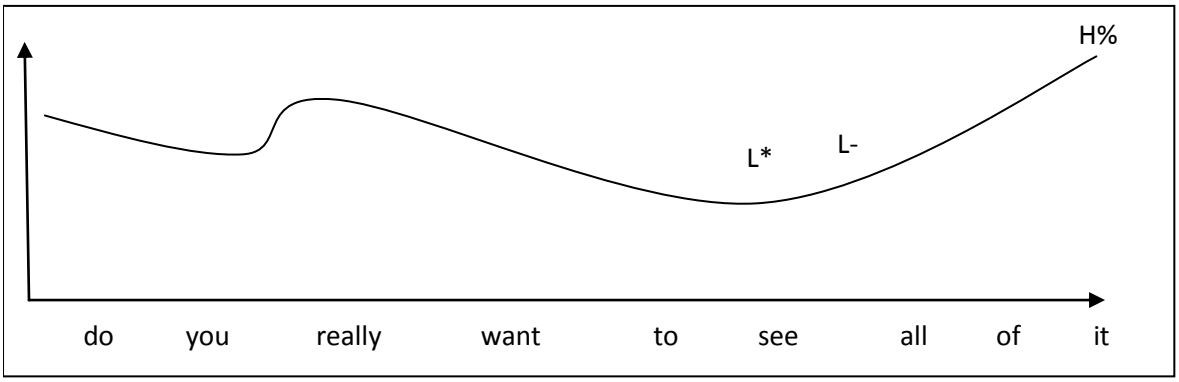
உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் ஒரு முக்கியமான வேலை மீக்கூறின் பொருத்தமான மொழி உருப்படுத்தத்தை உருவாக்குவாதாகும். அவற்றிலிருந்து வெளியீட்டு பேச்சு அலைவடிவில் வெளிப்படுத்தப்படும் பொருத்தமான ஒலியிக்க அமைப்பொழுங்கை உருவாக்குவதாகும். இத்தகைய மீக்கூறுப் பகுதியைக்கொண்ட ஒரு உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் வெளியீடு ஒரு கால அளவும் F0 (இசைமை) மதிப்பும் உள்ள ஒலிகளின் தொடச்சியாகும். ஒவ்வொரு ஒலியின் கால அளவுவும் ஒலியியல் சார் சூழலைப் பொறுத்து அமையும். F0 மதிப்பு சொல்சார் அசை அழுத்தம், வாக்கியத்தில் ஒலியழுத்தம் செய்யப்பட்ட அல்லது கவனக்குவிப்பு செய்யப்பட்ட கூறு, கூற்றின் இசையோட்டச் சுரம் இவற்றையும் உள்ளடக்கிய முன்னர் விளக்கிய காரணிகளால் பாதிக்கப்படுள்ளது. பின்வரும் படம் FESTIVAL-இலிருந்து (Black et al, 1999) Do you really want to see all of it? என்ற வாக்கியத்தின் உரையிலிருந்து பேச்சு ஒழுங்குமுறையின் மாதிரி வெளியீட்டைக் காட்டும். படம் இரண்டில் காட்டப்பட்ட F0 மதிப்புகளுடன் கூடிய இந்த வெளியீடு அலைவடிவ உருவாக்க கூறுக்கான உள்ளீடாகும். கால அளவு CART-நடை தீர்மானக் கிளையமைப்பால் கணக்கிடப்பட்டுள்ளது (Jurafsky and Martin, 2000:133).

படம் 1: Do you really want to see all of it? என்ற வாக்கியத்தின் FESTIVAL (Black et al., 1999) வெளியீடு. படம் 2-இல் சரியான இசையோட்ட வடிவத்தோற்றம் தர்ப்பட்டுள்ளது Jurafsky and Martin, 2000:133).

H* L* L-H%

do		You		really				want			to		see		all		of		it	
d	uw	y	u	r	lh	l	iy	w	a	n	t	t	a	s	iy	a	l	a	v	ih	t
			w						a				x			o		h			
11	11	5	5	7	6	5	8	5	5	7	4	4	4	5	13	7	9	4	6	4	12
0	0	0	0	5	4	7	2	7	0	2	1	3	7	4	0	6	0	4	2	6	0

மாதிரி வக்கியத்திற்கு FESTIVAL-ஆல் இணைப்பாக்க ஒழுங்கமைப்பால் உருவாக்கப்பட்ட F0 வடிவத்தோற்றம் (Jurafsky and Martin, 2000:133).

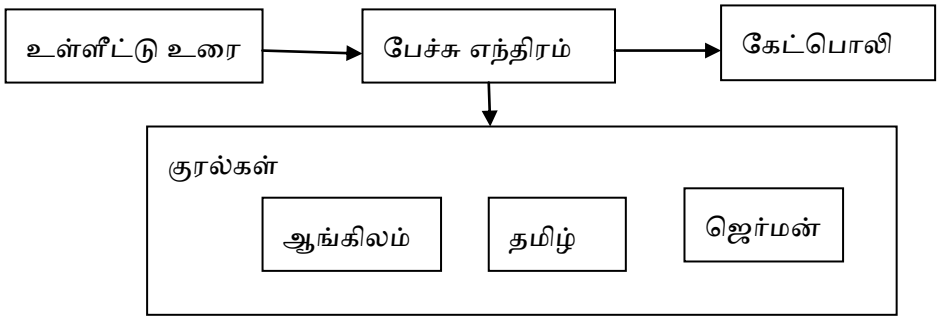


முன்னர் பரிந்துரைக்கப்பட்டது போல், எந்த அசையை ஒலியழுத்தம் செய்யவேண்டும் எந்த அசைக்குச் சுரத்தைப் பயன்படுத்தவேண்டும் போன்ற உண்மை உலக அறிவு மற்றும் பொருண்மையியல் சார் தகவல் தேவை ஆகையால் மீக்கூறு அமைப்பொழுங்கை தீர்மானிப்பது கடினமானதாகும். இவ்வகையிலான தகவல்களை உரையிலிருந்து பிரித்தெடுப்பது கடினம் எனவே மீக்கூறு தொகுதிகள் உள்ளீடு செய்யப்பட்ட பனுவலில் ஒரு "நடுநிலை அறிவிப்பு" பதிப்பை விளைவிக்க நோக்கம் கொண்டுள்ளது.

3.5.7. **உரையிலிருந்து பேச்சுக் கூட்டிணைப்பாக்கம்**

பேச்சுக் கூட்டிணைப்பாக்கம் மனிதப் பேச்சின் செயற்கைத்தன்மையான உற்பத்தி; இது படத்தில் காட்டியுள்ளது போன்று பேச்சு குறிகையை எழுத்துவடிவப் பனுவலாக மாற்றும் செயல்பாடாகும். மேலும் நாம் தமிழ் பனுவல்-பேச்சு ஒழுங்குமுறை உருவாக்குவதாய் இருந்தால் இவ்வொழுங்குமுறை எந்த எழுத்துவடிவ உள்ளீட்டையும் ஏற்கவேண்டும். இதைக் கருத்திற்கொண்டு கணினி பேச்சைப் பற்றி ஒரு சில வேறுபடுத்தல்கள் செய்வது தேவையாகும். ஏதாவது பேச்சை ஒலிப்பதிவு செய்து கணிப்பொறியில் சேமித்துவைத்து ஒலிக்கச்செய்து

கேட்கலாம். பல ஒழுங்குமுறைகள் இதைத்தான் செய்கின்றன; விடை எந்திரம் (answer machine) பதிவு செய்யப்பட்ட செய்திகளைத் திரும்ப ஒலிக்கின்றது; வானொலிப்பெட்டி ஏற்கனவே பதிவுசெய்யப்பட்ட நேர்காணலைத் திரும்ப ஒலிக்கின்றது; இவ்வாறு பல ஒழுங்குமுறைகள் உள்ளன. உரை-பேச்சுக்குப் பின்னணியாக வரும் கருத்து என்னவென்றால் "திரும்ப ஒலிக்கப்படும்" செய்திகள் உண்மையிலேயே ஒலிப்பதிவு செய்யப்பட்டவை அல்ல. திரும்ப ஒலிக்கச் செய்யும் நடத்தையிலிருந்து அடுத்தபடியான நடவடிக்கை பொதுவான சொற்களை ஒலிப்பதிவு செய்து திரும்பச் சேர்ப்பதாகும்; இந்த நுட்பம் தொலைபேசி உரையாடல் சேவைகளில் அடிக்கடி பயன்படுத்தப்படுகின்றது. சில வேளைகளில் பயன்விளைவு ஏற்றுக்கொள்ளக்கூடியதாய் இருக்கும்; சில வேளைகளில் ஏற்றுக்கொள்ள இயலாததாய் இருக்கும்; செயற்கையாக ஒன்றுசேர்க்கப்பட்டப் பேச்சு செயற்கையாக இருக்கும். மாறாக உரை-பேச்சு விரும்பப்படுகிற செய்தி உண்மையிலேயே பேசப்பட்டாலும் இல்லாவிட்டாலும் எதையும் பேச்சும் இலக்கைக் கொண்டது.



3.5.7.1. பேச்சு

பேச்சுக் குறிகை (speech signal) ஒலி மூலத்திற்கும் கேட்பவருக்கும் இடையிலுள்ள ஊடகத்தில் ஏற்படும் அழுத்த மாற்றங்களின் வரிசையாகும். பேச்சுக் குறிகையின் மிகப் பொதுவான உருப்படுத்தம் பெரும்பாலும் அலைவடிவம் (wave form) என்று அழைக்கப்படும் ஆசிலோகிராம் (ocillogram) ஆகும். பேச்சு மனித முன் தொண்டை, தொண்டை, வாய், நாக்கு இவற்றின் வழியாகச் செல்லும் காற்றின் போக்கில் ஏற்படும் தடைகளால் உற்பத்தி செய்யப்படுகின்றது. உயிரொலிகளை உள்ளடக்கிய குரல் ஒலிகள் ஓரளவு சீரான அமைப்பொழுங்கைக் காட்டுகின்றன. குரலொலிகள் இயல்பாக அதிக சக்தியைக் கொண்டிருக்கின்றது; *ஸ்* போன்ற குரலிலா ஒலிகள் சீரற்ற அமைப்பொழுங்கைக் கொண்டிருக்கும். ஒரு பேச்சு குறிகையை ஒலியன்களின் அல்லது நோட்டங்களின் வரிசையாகப்

புரிந்துகொள்ளப்படலாம். பொதுவாகக் கூற்று அதிக அளவில் வழக்கமாகப் பயன்படுத்தப்படுகின்ற ஐ.பி.எ. ஒலியியல் நெடுங்கணக்கால் (IPA Phoneic Alphabet) ஒலிபெயர்க்கப்படுகின்றது.

பேச்சு குறிகையை உருப்படுத்தம் செய்யும் மற்றொருவழி இசைமை ஆய்வால் உற்பத்திசெய்யபடும் ஒன்று. பேச்சு இரு பாகங்களைக் கொண்ட இயற்பியல் செயற்பாங்கு ஆகும்: ஒலி மூலத்தின் (குரல்வளை மடல்கள்) மற்றும் வடிகட்டுதலின் (நாக்கு, உதடுகள், பல் போன்றவற்றால்) உற்பத்திப் பொருள். இசைமை ஆய்வு இறுதி பேச்சுக் கூற்றை ஆய்வதால் ஒலி மூலத்தின் அடிப்படை நிகழ்வெண்ணைக் கண்டுபிடிக்க முயலுகின்றது. அடிப்படை நிகழ்வெண் குரல்வளை மடல்களால் உற்பத்தி செய்யப்படும் ஒலியின் முனைப்பான நிகழ்வெண்ணாகும். எவ்வாறு கேட்பவர்கள் பேசுபவர்களின் இசையோட்டதையும் அசை அழுத்தத்தையும் கேட்கின்றனர் என்பதற்கு வலுவான தொடர்புப்பொருத்தம் அடிப்படை நிகழ்வெண்ணாகும். ஆணின் இயல்பு F0 நிகழ்வெல்லை 80-200 Hz ஆகும்; பெண்களுக்கு 150-350 Hz ஆகும்.

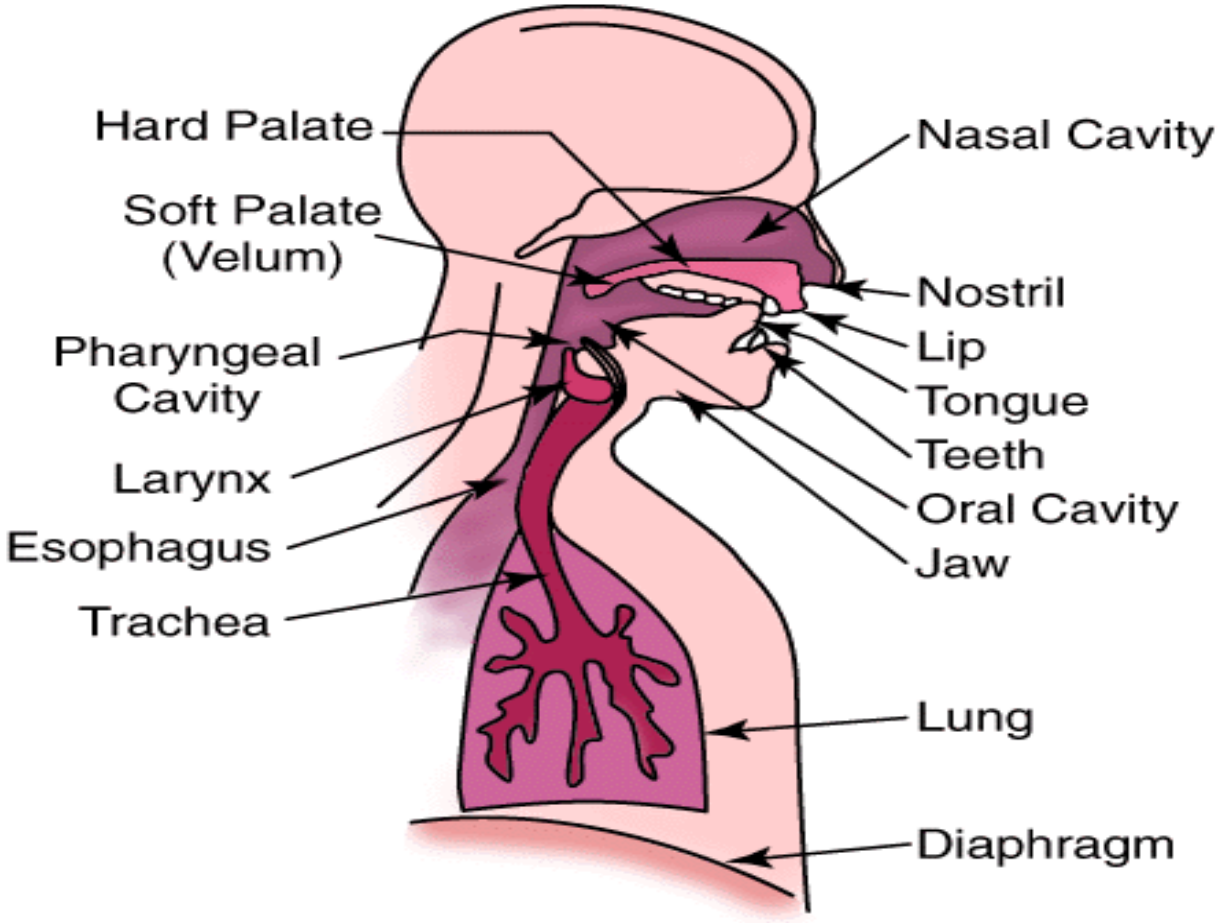
3.5.7.2 பேச்சு உற்பத்தி

பேச்சு நுரையீரலிலிருந்து குரல்வளை, வாயறை, மூக்கறை வழியாகச் செல்லும் காற்றொழுக்கால் உருவாக்கப்படுகின்றது. இது மூன்று செயற்பாங்குகளை உட்படுத்தும்: தொடக்கம் (initiation), இழுமை (phonation), வாய்-மூக்குச் செயற்பாங்கு (oral-nasal process), ஒலிப்பு (articulation). நுரையீரல்களிலிருந்து காற்று வெளியேற்றப்படுவது தொடக்கச் செயற்பாங்கு ஆகும். தொண்டையில் நிகழும் செயற்பாங்கு இழுமை ஆகும்.

தொண்டைக் காற்று வழியில் இரண்டு செங்குத்தான மடல்களைக் கொண்டிருக்கின்றது; அவை குரல்வளை மடல்களாகும் (vocal folds). மடல்களுக்கு இடையிலுள்ள இடைவெளி குரல்வளையாகும் (glottis). குரல்வளையை அதன்வழியே காற்று போகாதபடி மூட இயலும் அல்லது குரல் ஒலிகளை ஒலிக்கும் விதத்தில் குரல்வளை மடல்களை அதிரச்செய்ய இயலும் படி குறுகிய திறப்பைக் கொண்டிருக்கச் செய்யவியலும். இறுதியாக இயல்பான மூச்சுவிடும் போது அதை அதிகமாகத் திறக்க இயலும்; இதனால் குரல்வளை மடல்களின் அதிர்வு குறைக்கப்பட்டு குரலிலா ஒலிகள் உற்பத்தி செய்யப்படும்.

தொண்டை, மேல்தொண்டை வழியாகச் சென்ற பின்னர் காற்று வாயறை அல்லது மூக்கறை வழியாகச் செல்ல இயலும். பின்னண்ணம்/கடையண்ணம் இந்தத் தேர்வுக்குப் பொறுப்பான பகுதியாகும். வாய்-மூக்குச் செயற்பாங்கு மூக்கு மெய்யொலிகள் மற்றும் பிற ஒலிகளைத் தீர்மானிப்பதற்கு உதவும். இறுதிச் செயற்பாங்கு வாயறையில் நிகழும் ஒலிப்புச்

செயற்பாங்கு ஆகும்; இந்தச் செயற்பாங்கு பெரும்பாலான பேச்சொலிகளை வேறுபடுத்த மனிதனுக்கு உதவுகின்றது. பேச்சு உற்பத்தியின் உடலியங்கியல்/உடற்கூறியல் கீழே படத்தில் தரப்பட்டுள்ளது. பேச்சு உற்பத்தி பற்றி இயல் 3இல் விரிவாகக் கூறப்பட்டுள்ளது.



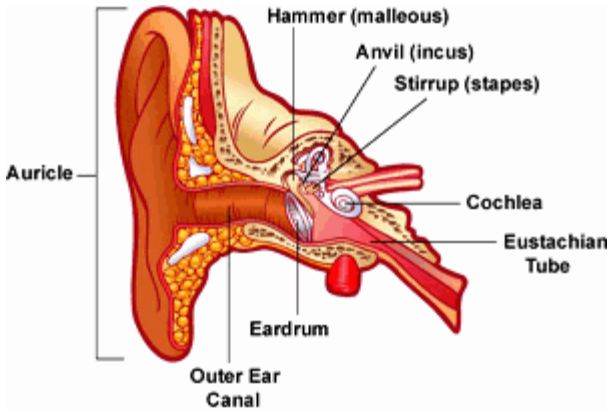
Hard palate 'வல்லண்ணம்', trachea 'முச்சுகுழல், teeth 'பல், Soft palate 'மெல்லண்ணம்', nasal cavity 'மூக்கறை, oral cavity 'வாயறை' Pharangeal cavity 'மேல்தொண்டைஅறை, nostril 'மூக்குத் துளை, jaw 'தாடை, Larynx 'தொண்டை' lip 'இதழ்', lung 'நுரையீரல்', Esophagus 'உணவுகுழல்', tongue 'நா', diapharam 'உந்தி/உதரவிதானம்'.

3.5.7.3 பேச்சு கேட்புணர்வு

கேட்புணர்வு ஒழுங்கமைப்பில் இரண்டு முக்கியமான கூறுகள்/பகுதிகள் உள்ளன: வெளிப்படையான கேட்பு உறுப்புகள் (காதுகள்) மற்றும் கேட்பு நரம்பு ஒழுங்குமுறை (மூளை). காது ஒலிப்பு அழுத்த குறிகையை ஆய்கின்றது; முதலில் அதைப் பேசிலியார் தோல்/ஜவ்வு மீது எந்திரத்தன்மையான அதிர்வு அமைப்பொழுங்குகளாக மாற்றுகின்றது; பின்னர் கேட்பு நரம்பால்

கடத்தப்பட்ட துடிப்புகளின் வரிசைகளால் உருப்படுத்தம் செய்கின்றது. கேட்புணர்வுத் தகவல் கேட்பு நரம்பு ஒழுங்கமைப்பின் பல்வேறு நிலைகளில் பிரித்தெடுக்கப்படுகின்றது.

மனிதக் காது மூன்று பகுதிகளைக் கொண்டது: வெளிக்காது, நடுக்காது, அகக்காது. வெளிகாது காணக்கூடிய வெளிப்பகுதி, ஒலி பயணிக்கும் குழாய் வடிவ வெளி கேட்புக் குழாய் (external auditory canal) இவற்றைக் கொண்டிருக்கும். குழாய் அதன் ஒரு முனையில் காதுமுரசால் (ear drum) மூடப்பட்டிருக்கும். காற்று அழுத்த வேறுபாடுகள் காதுமுரசை அடைந்தது அதை அதிரச்செய்யும்; இந்த அதிர்வுகளை அதன் எதிர் பக்கம் இருக்கும் எழும்புகளுக்குக் கடத்தும். உள்வரும் ஒலி அழுத்த அலை போன்ற அதே நிகழ்வெண்ணில் (சுருங்குதல், விரிதல் இவை மாறிமாறி நிகழ்தல்) காதுமுரசின் அதிர்வெண் இருக்கும். நடுக்காது காற்று நிரப்பப்பட்ட இடைவெளி அல்லது அறை ஆகும். இந்த அறையுடன் மூக்கையும் தொண்டையையும் இணைக்கும் குழாய்வழி காற்று நடுக்காது அறைக்குப் பயணிக்கும். அகக் காதுக்கு (காக்ளியா) (cochlea) எலும்பு இடைமுகத்தில் இருக்கும் சிறிய தோல் ஓ வடிவச் சாளரம் (oval window) ஆகும். காக்ளியாவின் சுவர்கள் எலும்பாக இருப்பதால் சக்தியானது ஸ்டேப்களின்/உட்செவி எலும்புகளின் (stapes) எந்திரம்சார் செயல்பாட்டல் ஓ-வடிவ சாளரத்தின் மேல் போர்த்தப்பட்டுள்ள/இழுக்கப்பட்டுள்ள தோல்/ஜவ்வுமீது ஒரு பதிப்பாகக் கடத்தப்படுகின்றது. காதின் அமைப்பு படத்தில் காட்டப்பட்டுள்ளது.



Hammer ‘சுத்தி’, Auricle ‘காதுமடல்’, Outer ear canal ‘வெளிக்காது குழாய், Anvil ‘பட்டறை எலும்பு’, Stirrup ‘அங்கவடி எலும்பு, Cochlea ‘அகக்காது, Ear drum ‘காது முரசு, Eustachian tube ‘நடுச்செவிக் குழல்’

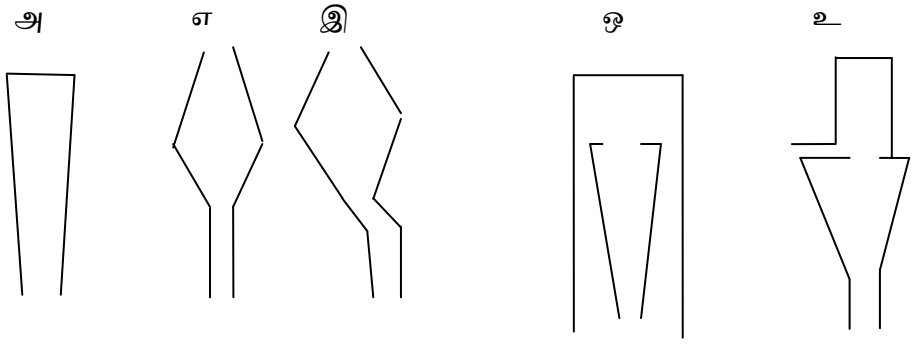
ஒலி கேட்புணர்வுக்கு அகக்காதின்/உட்காதின் தேவையான அமைப்பு/பகுதி காக்ளியா ஆகும்; இது ஒலியின் உருப்படுத்தத்தை மூளைக்கு கடத்தும் கேட்பு நரம்புடன் நேரடியாகத் தொடர்புகொள்கின்றது. சுருள் பகுக்கப்பட்டுள்ளது; முதன்மையாக நீளவாக்கில் ஓடும் பாசிலார்

ஜவ்வாலும்/தோலாலும் இரண்டு திரவம் நிரப்பப்பட்ட அறைகளாலும் பகுக்கப்பட்டுள்ளது. காக்ளியாவை ஏறக்குறைய வடிகட்டும் வங்கியாயக் கருத இயலும்; இதன் வெளியீடுகள் அதிர்வெண்-இட மாற்றம் நிறைவேற்றப்பட வேண்டி இடத்தால் ஒழுங்குபடுத்தப்பட்டுள்ளது. காக்ளியாவின் அடிக்கு அண்மையிலிருக்கும் வடிகட்டிகள் உயர்ந்த நிகழ்வெண்களுக்கு பதிலளிப்பு செய்கின்றது; உச்சிக்கு அண்மையில் இருப்பவைகள் குறைந்த நிகழ்வெண்களுக்குப் பதிலளிப்பு செய்கின்றது. நரம்புக் கடத்தல் செயற்பாங்கு இதைத் தொடர்கின்றது மற்றும் ஸ்பெக்ட்ரல் குறிகையைக் (spectral signal) கிட்டத்தட்ட பண்புக்கூறு பிரித்தெடுக்கும் கூறு/பகுதிக்கு பொருத்தமுறும் கேட்பு நரம்பின் (auditory nerve) செயல்பாட்டு குறிகைகளாக மாற்றுகின்றது.

3.5.7.4 பேச்சுக் கூட்டிணைப்பாக்கத்தின் வரலாறு

பல நூற்றாண்டுகளாக மனித இனத்தின் கனவாகச் செயற்கைப் பேச்சு இருந்தது. குறிகை ஆய்வைக் கண்டுபிடிப்பதற்கு முன்னர் மனிதப் பேச்சை உருவாக்க இயலும் சில எந்திரங்களை உருவாக்குவதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டன. பேச்சுக் கூட்டிணைப்பாக்கம் (speech synthesis) இயந்திரச்சார்பிலிருந்து மின்சாரம்சார் கூட்டிணைப்பாக மாறியது; இயந்திரம்சார் கூட்டிணைப்பாக்கம் 1779-இல் தொடங்கப்பட்டது. 1779-இல் ரஷ்யப் பேராசிரியரான கிறிஸ்டியன் கிராட்செண்டைன் (Christian Kratzenstein) ஐந்து நெட்டுயிர்களுக்கு (அ, எ, இ, ஒ) இடையேயுள்ள உறுப்பியல்சார் வேறுபாடுகளை விளக்கினார்; அவர் அவற்றைச் செயற்கையாக உற்பத்திசெய்யக் கருவியை உருவாக்கினார். கிராட்செண்டைன் படத்தில்காட்டியுள்ளது போன்று மனிதப் பேச்சுக்குழாயுடன் ஒற்றுமையுள்ள ஒலியியக்க ஒலியூக்கிகளை (acoustic resonators) உருவாக்கினார்; இசைக்கருவிகள் போன்று அதிர்வுறும் நாக்கால் ஒலியூக்கிகளைச் செயற்படுத்தினார்.

கிராட்செண்டைனின் ஒலியூக்கிகள் (Kratzenstein’s resonators)

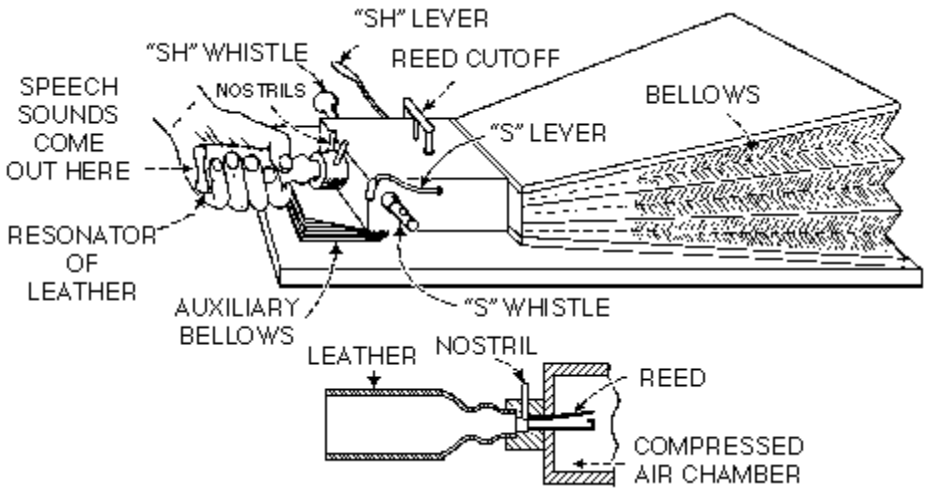


1791-இல் வோல்ஃப்காங்க் வோன் கெம்ப்ளென் (Wolfgang von Kempelen) “ஒலியியக்க எந்திரப் பேச்சு எந்திரம்” (Acoustic-Mechanical Speech Machine) என்பதை அறிமுகப்படுத்தினார். அது சில தனி ஒலிகளையும் ஒலிச் சேர்க்கைகளையும் உற்பத்தி செய்ய முடிந்தது.

வோல்ஃப்காங்க் வோன் கெம்ப்ளென் ஒலியியக்க எந்திர பேச்சு எந்திரம்

அதன்பின்னர் 1837-இல் சார்லஸ் வீட்ஸ்டோன் (Charles Wheatstone) என்பவர் வான் கெம்ப்ளெனின் மாதிரி அடிப்படையில் ஒரு பேசும் எந்திரத்தை உருவாக்கினார். இந்தப் பேச்சு எந்திர வகை உயிரொலிகளையும் பெரும்பாலான மெய்யொலிகளையும் உற்பத்திசெய்ய முடிந்தது; இது ஒலிச்சேர்க்கைகளையும் முழுச் சொற்களையும் உற்பத்தி செய்ய முடிந்தது.

வீட்ஸ்டோனின் பேசும் எந்திரம் (Wheatstone's speaking machine)

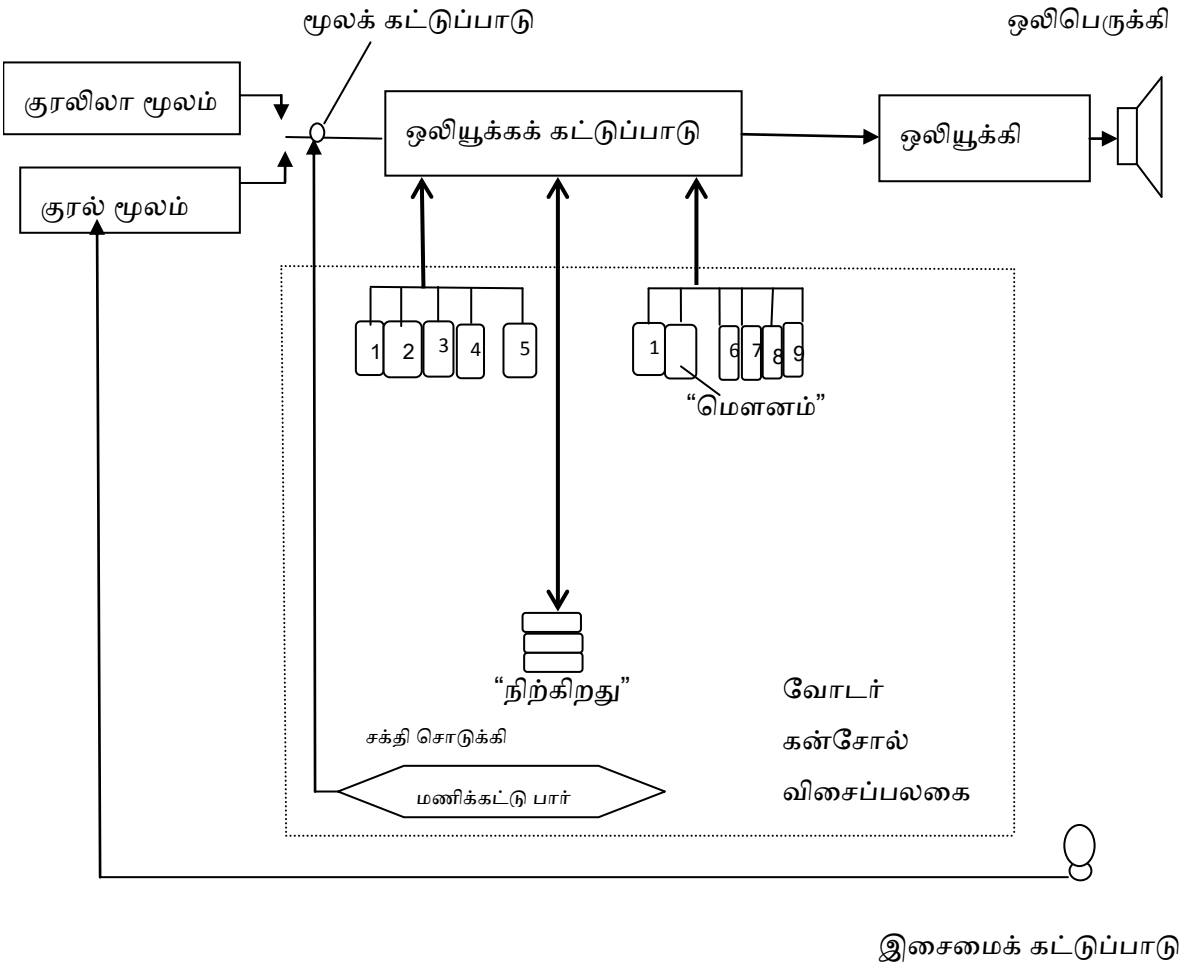


மின்னியல்சார் கூட்டிணைப்பாக்கக் கருவி (Electrical synthesis device) 1922-இல் உருவாக்கப்பட்டது. முதல் மின்னியல்சார் கூட்டிணைப்புக் கருவி ஸ்டேவார்ட் (Stewart) என்பவரால் அறிமுகப்படுத்தப்பட்டது. இந்த எந்திரம் தனி நிலையான உயிரொலிகளை உற்பத்தி செய்தது; ஆனால் மெய்யொலிகளையும் கூற்றுக்களையும் உற்பத்தி செய்யவில்லை. 1932-இல் ஒபாட்டா (Obata), டெஷிமா (Teshima) என்ற ஜப்பனிய ஆய்வாளர்கள் உயிரொலிகளில் மூன்றாவது ஒலிச்செறிவைக் கண்டுபிடித்தனர். முதல் மூன்று ஒலிசெறிவுகளும் பொதுவாகப் புரியத்தக்கப் பேச்சு கூட்டிணைப்பாக்கத்திற்குப் போதுமானதாகக் கருதப்பட்டது.

இருப்பினும் பேச்சுக் கூட்டிணைப்பானாகக் (speech synthesizer) கருத்தப்பட்ட முதல் கருவி 1939-இல் ஹோமர் டட்லே (Homer Dudley) என்பவரால் அறிமுகப்படுத்தப்பட்டது. இந்தக் கருவி வோடர் (VODER – Voice Operating Demonstrator) என்று பெயரிடப்பட்டது. இது படத்தில் காட்டப்பட்டுள்ளது. இது விசைப்பலகையால் இயக்கப்படும் மின்னணுசார் பேச்சு ஆய்வியாகும்; கூட்டிணைப்பான் 1930களின் மையத்தில் பெல் லாபால் (Bell lab) உருவாக்கப்பட்டது. வோடரைப் பயன்படுத்துவது கடினம். மட்டுமன்றி இந்த இயந்திரத்தால் உற்பத்திசெய்யப்பட்ட பேச்சின் பண்பும் புரிதிறனும் நல்லதல்ல. இந்த ஒழுங்குமுறை செயற்கைப் பேச்சை உற்பத்தி செய்வதில் அதன் திறன் காரணமாகப் புகழ்பெற்றது.

Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

வோடர் பேச்சுக் கூட்டிணைப்பான் (VODER speech synthesizer)



வோடரின் செய்முறை விளக்கத்திற்குப் பின்னர் அறிவியல் உலகம் பேச்சு கூட்டிணைப்பாக்கத்தில் அதிக விருப்பத்தைக் காட்டியது. வோடரின் அடிப்படை அமைப்பும் கருத்தும் பேச்சின் மூலம்-வடிகட்டுவான்–மாதிரி (source-filter-model) தற்போதைய ஒழுங்குமுறைகள் போன்றதே. பிராங்கிளின் கூப்பர் (Franklin Cooper) மற்றும் அவர்தம் உதவியாளர்கள் 1951-இல் ஹாஸ்கின்ஸ் ஆய்வுக்கூடத்தில் (Haskins laboratory) அமைப்பொழுங்கு திருப்பி ஒலிக்கும் கூட்டிணைப்பானை (pattern playback synthesizer) உருவாக்கினர். இது பதிவுசெய்யப்பட்ட நிறமாலை நிழற்பட/ஒலிவர்ணனைப்பட அமைப்பொழுங்கை மூல அல்லது மாற்றப்பட்ட வடிவில் ஒலிகளாகத் திரும்ப மாற்றியது. நிறமாலை நிழற்பட/ஒலிவர்ணனைப்பட அமைப்பொழுங்கு ஒளிப்புகவல்ல பெல்ட் மீது ஒளிசார்ந்ததாக/ஆப்டிக்கலாகப் பதிவிக்கப்படுகின்றது.

PAT (Pragmatic Artificial Talker) முதல் ஒலிச்செறிவு கூட்டிணைப்பாக்கி ஆகும்; இது வால்டர் லாரன்சால் (Walter Lawrence) 1953-இல் அறிமுகப்படுத்தப்பட்டது. பாட் இணையாக இணைக்கப்பட்ட மின்னணு ஒலிச்செறிவு ஒலியூக்கிகளைக் கொண்டிருக்கின்றது. உள்ளீட்டு குறிகை பஸ்-ஆகவோ இரைச்சலாகவோ (buzz or noise) இருக்கும். நகரக்கூடிய கண்ணாடி ஸ்லைடு (moving glass slide) மூன்று ஒலிச்செறிவு நிகழ்வெண்களை (formant frequencies), குரல், அலைவீச்சு, அடிப்படை அதிர்வெண் (fundamental frequency), இரைச்சல் அலைவீச்சு (noise amplitude) இவற்றைக் கட்டுப்படுத்த ஆறு காலச் செயல்பாடுகளாக வர்ணம்பூசிய அமைப்பொழுங்குகளை மாற்றப் பயன்படுத்தினர். குன்னார் ஃபண்ட் (Kunnar Fant) 1950-இல் முதலில் ஒழுக்கு ஒலிச்செறிவு கூட்டிணைபாக்கி (cascade formant synthesizer) ஓவ் 1 (OVE I) (OVE- Orator Verbis Elecris) அறிமுகப்படுத்தினார்; இது ஒழுக்கில் இணைக்கப்பட்ட ஒலிச்செறிவு ஒலியூக்கிகளைக் கொண்டது. 1962-இல் ஃபாண்டும் மார்ட்டோனியும் (Fant and Martony) மேம்படுத்தப்பட்ட ஓவ் 2 (OVE II) கூட்டிணைப்பாக்கியை அறிமுகப்படுத்தினார்; இது உயிரொலிகள், மூக்கொலிகள், தடைமெய்யொலிகள் இவற்றிற்குப் பேச்சுக்குழலின் கடத்தல் செயல்பாட்டை மாதிரிப்படுத்தத் தனிப் பகுதிகளைக் கொண்டிருக்கின்றது. குரல், உயிர்ப்பு இரைச்சல், உராய்வு இரைச்சல் என்பன சாத்தியமான துண்டல்கள் ஆகும்.

பாட் மற்றும் ஓவ் (PAT and OVE) இவற்றை ஆய்ந்த பின்னர் ஜாண் ஹோல்ம்ஸ் (John Holmes) 1972-இல் இணை ஒலிச்செறிவு கூட்டிணைப்பாக்கியை (paralle formant synthesizer) அறிமுகப்படுத்தினார். 1978-இல் ரிச்சர்ட் காக்னான் (Richard Gagnon) மலிவான வோட்ராக்ஸ் அடிப்படையிலான டைப்-என்–டாக் ஒழுங்குமுறையை (Votrax-based Type-n-Talak system) அறிமுகப்படுத்தினார். 1982-இல் ஸ்ட்ரீட் எலக்ட்ரானிக்ஸ் (Street Electronics) எதிரொலி மலிவு ஒலியிருமை கூட்டிணைப்பாக்கியை (Echo low-cost diaphone synthesizer) அறிமுகப்படுத்தினார். தற்காலப் பேச்சுக் கூட்டிணைப்பாக்கத் தொழில்நுட்பங்கள் மிகச் சிக்கலான நவீனமான நெறிமுறைகளையும் வழிமுறைவரைவுகளையும் உள்ளடக்கியது. ஒன்றிணைப்பாக்கப் பேச்சுக் கூட்டிணைப்பாக்கம், ஹிட்டன் மார்க்கோவ் மாதிரிகள் (HMM models) அடிப்படையிலான கூட்டிணைப்பாக்கம் என்பன அண்மையில் பயன்படுத்தப்பட்டுள்ள தற்கால நெறிமுறைகள் ஆகும்.

3.5.7.5. பேச்சுக் கூட்டிணைப்பாக அணுகுமுறைகள்

பேச்சு மனிதவினம் ஒருவருக்கொருவர் கருத்துக்களைப் பரிமாறிக்கொள்ளும் இயற்கையான முறையாகும். பேச்சுக் கூட்டிணைப்பாக்கம் செய்கையாக, குறிப்பாகக்

கணினியைப் பயன்படுத்திப் பேச்சைக் கூட்டிணைப்பாக்கம் செய்யும் செயல்முறையாகும். மனிதப் பேச்சை பாவிக்கும்/போலச்செய்யும் ஒலி உருவாக்கம் கீழ் மட்டக் கூட்டிணைப்பாக்கம் என்று கூறப்படுகின்றது. மேல்மட்ட கூட்டிணைப்பாக்கம் எழுதப்பட்ட உரையை/பனுவலை அல்லது குறியீடுகளைக் கீழ்மட்ட உருவாக்க ஒழுங்குமுறையை இயக்கப் பொருத்தமான விருப்பப்பட்ட ஒலியியக்கக் குறியீடுகளின் அருவ உருப்படுத்தமாக மாற்றுவதை விளக்குகின்றது. உற்பத்தியின் நெறிமுறைகளின் அடிப்படையில் பேச்சு கூட்டிணைப்பாக்கத்தை நான்கு முக்கியமான வகைப்பாடுகளாக வகைப்படுத்தலாம்.

1 ஒலிப்பியல் கூட்டிணைப்பாக்கம் (Articulatory synthesis)

2. ஒலிச்செறிவு கூட்டிணைப்பாக்கம் (Forman snynthesis)

3. இணைப்பு கூட்டிணைப்பாக்கம் (Concatenative synthesis)

4. எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பாக்கம் (HMM based synthesis)

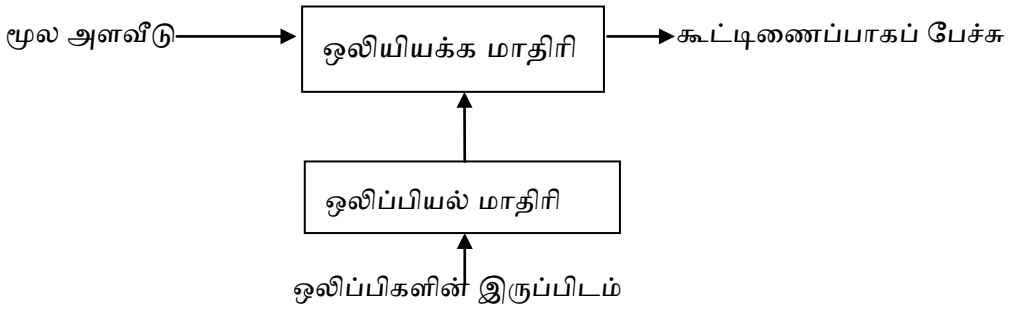
3.5.7.5.1. ஒலிப்பியல்சார் கூட்டிணைப்பாக்கம்

பேச்சுக் குழாயின் மாதிரியைப் பயன்படுத்தி பேச்சு ஒலிகளை உற்பத்திசெய்வது ஒலிப்பியல் சார் கூட்டிணைப்பாக்கம் (articulatory synthesis) ஆகும்; இது நேரடியாகவோ மறைமுகமாகவோ பேச்சு ஒலிப்பிகளின் இயக்கத்தை ஊக்குவிக்கின்றது. பேச்சுக்குழாயின் வடிவத்தைப் பல வழிகளில் கட்டுப்படுத்த இயலும்; இது நாக்கு, தாடை, உதடுகள் போன்ற பேச்சு ஒலிப்பிகளின் இருப்பை மாற்றுவதை உட்படுத்தும். பேசுக்குழாயின் உருப்படுத்தம் வழி காற்றின் ஓட்டத்தை டிஜிட்டலாக போலச்செய்வதால்/பாவிப்பதால் பேச்சு உருவாக்கப் படுகின்றது. ஒலிப்பியல்சார் கூட்டிணைப்பாகத்தில் இணை ஒலிப்பு விளைவுகள் இயற்கையாக வருகின்றது. குரல்வளைசார் மூலப் பண்புக்கூறுகள், பேச்சுக்குழாய் மற்றும் குரல்வளை மடல்கள் இவற்றிற்கிடையேயுள்ள ஊடாட்டம், துணை குரல்வளைசார் ஒழுங்குமுறையின் பங்களிப்பு மற்றும் மூக்குக் குழாய் மற்றும் மூக்குத் துளைகள் (sinus cavities) இவற்றுடன் இதைச் சரியாகக் கையாள இயலும்.

ஒலிப்பியல்சார் உருவாக்கம் படத்தில் காட்டியுள்ளது போன்று இரு மதிரிகளை இணைக்கின்றது. ஒலிப்பியல் மாதிரியில் பேச்சுக் குழல் பல சிறிய பகுதிகளாகப் பகுக்கப்பட்டுள்ளது மற்றும் தொடர்புடைய குறுக்கு வெட்டுப் பகுதிகள் (cross-sectional areas) பேச்சுக் குழாய்ப் சிறப்புப் பண்புகளை உருப்படுத்தம் செய்வதற்கு அளவீடுகளாகப் பயன்படுத்தப்படுகின்றது. ஒலியியக்கவியல் மாதிரியில் ஒவ்வொரு குறுக்கீடான பரப்புகளும்

மின்சார அனலாக் கடத்து லைனால் (electrical analog transmission line) தோராயமாக்கப்பட்டுள்ளது.

கீழே தரப்பட்டுள்ள ஒலிப்பியல் சார் மாதிரி பேச்சு உற்பத்தியின் முக்கியமான கூறுகளை உருப்படுத்தம் செய்கின்றது; இது தனது எல்லைகளை முறையே குரல்வளையிலும் வாயிலும் இடமாகக் கொண்ட ஒலியியக்கக் குழாயின் குறுக்குவெட்டுப் பரப்பின் வேறுபாட்டைப் பிரதிபலிக்கும் பரப்புச் செயல்பாட்டின் கணிப்பு அதன்முக்கிய நோக்கமாகும்.



பேச்சுக்குழாயின் ஜியாமண்டரிக் பரப்பு (geomenteric area) ஒலிப்பைப் பற்றிய நமது புரிந்துகொள்ளலுக்கு முக்கியமாகும் மற்றும் அது பேச்சு உருவாக்கத்திற்கு முக்கியக் காரணியாகும். பேச்சு உருவாக்கத்தின் ஒலியியக்கக் கோட்பாடு அடிப்படையில் சீரற்ற மற்றும் கால மாறுபாட்டுடைய குறுக்குவெட்டுப் பரப்புகள் மனிதப் பேச்சுக் குழாயை ஒரு ஒலியக்கக் குழாயாக (acoustic tube) மாதிரியாக்கம் செய்ய இயலும். வேறுபட்ட மொழி ஒலிகளை உற்பத்தி செய்ய அது கிளர்ச்சி மூலத்தை (excitation source) ஒழுங்குபடுத்துகிறது. ஒலிப்பு இடங்கள் குறிப்பிடப்பட்டுவிட்டால் குறுக்குவெட்டுப் பரப்புகள் பேச்சுக்குழாய் எல்லைக்கோட்டின் மீது ஒரு சல்லடை அமைப்பை (grid structure) ஒன்றின் மேல் ஒன்றாக வைப்பதால் கணிக்கப்படுகின்றது. இந்த சல்லடைக் கோடுகள் (grid lines) ஒலிப்பிகளின் இருப்பிடம் அடிப்படையில் வேறுபடுகின்றது (அவைகள் மெர்மெல்ஸ்டென் மாதிரியில் (Mermelstein's model) நிறுவப்படுகின்றது). மொத்தம் 60 பகுதிகள், பேச்சுகுழாய்க்கு 59 பகுதிகள் மற்றும் நமது மாதிரியில் பயன்படுத்தப்படும் குரல்வளையின் வெளியேற்றும் வழிக்கு ஒரு பகுதி (நிலையான நீளம் மற்றும் பரப்பு) நமது மாதிரியில் பயன்படுத்தப்படுகின்றது. அம்புவடிவ தூரம் (sagittal distance) பின்வய-மேல்வய மற்றும் முன்வய-கீழ்வய வெளிஎல்லைக் கோடுகளுக்கு இடையேயுள்ள சல்லடைக் கோடுப் பகுதியாக வரையறை விளக்கம் செய்யப்படுகின்றது.

பேச்சுக்குழாயின் மையக்கோடு பக்கத்திலுள்ள சல்லடைக் கோடுகளின் மையப் புள்ளிகளை இணைப்பதால் உருவாக்கப்படுகின்றது. மையக் கோட்டின் நீளம் பேச்சுக்குழாயின் நீளத்திற்குச் சமமானதாகக் கருதப்படுகின்றது. அம்புவடிவ தூரம் இறுதியில் அனுபவாத சூத்திரங்களால் (empiric formulas) குறுக்குவெட்டு பரப்புகளாக மாற்றப்படுகின்றது. பேச்சு உற்பத்தியின் ஒலியியக்கக் கோட்பாடு அடிப்படையில் நாம் தரப்பட்ட பேச்சுக்குழல் குறுக்குவெட்டுப் பரப்பிலிருந்து ஒலிச்செறிவு நிகழ்வெண்களைக் கணிக்க இயலும். தரப்பட்ட பேச்சுக்குழலின் ஒலியியக்கவியல் மாற்றச் செயல்பாட்டைக் கணிப்பதிலிருந்து ஒலிச்செறிவு நிகழ்வெண்களை ஆக்கக்கூறுகளாகப் பிரிப்பது எளிது. பேச்சுக்குழல் குறுக்குவெட்டுப் பரப்பு பேச்சுக் குறிகளைப் பற்றிய ஒலியியக்கவியல் தகவல்களைத் (ஒலிச்செறிவு நிகழ்வெண்களை) தருகின்றது. இறுதியாக நாம் அதைக் கட்டுப்பாடுகளுடன் கூடிய பல்பரிமாண வரிசையல்லாத உத்தமமாக்கச் சிக்கலாகக் (multi dimensional nonlinear optimization problem) கருதுகின்றோம் மற்றும் உத்தம ஒலிப்பு வெக்டாராகக் (optimum articulatory vector) கணிக்கின்றோம்.

மனிதப் பேச்சு ஒழுங்கமைப்பின் ஒலியியக்க மாதிரி பல துணைமாதிரிகளைக் கொண்டுள்ளது. பேச்சுக்குழல் மற்றும் மூக்குக்குழல் மாதிரிகள் (vocal tract and nasal tract models) இந்தக் குழல்களில் ஓசை பரவலைப் பாவிக்கின்றது. கிளர்வுறல் மூல மாதிரி பேச்சுக்குழலுக்கு குரல் கிளர்வுறல் அலைவடிவுகளை (voice excitation waveforms) உருவாக்குகின்றது. உரசொலிகள் மற்றும் அடைப்பொலிகளுக்குரிய இடுக்குகளில் கிளர்ந்தெழும் காற்றொழுக்கு ஓசை மூலத்தால் உருவாக்கப்படுகின்றது. பரவல் மாதிரி (radiation model) உதடுகளிலிருந்தும் மூக்குத்துவாரங்களிலிருந்தும் பரவும் ஒலியியக்கச் சக்தியைப் பாவிக்கின்றது. பேச்சுக்குழல், எலும்பு உட்புழைகளுடன் கூடிய மூக்குக் குழல், குரல்வளை தடை, துணைக் குரல்வளை குழல், கிளர்தல் மூலம், மற்றும் கிளர்தல் இரைச்சல் மூலம் இவற்றை உட்படுத்தும் பேச்சு ஒழுங்குமுறையின் கடத்தல்-கோடு சுற்று மாதிரி (transmission-line circuit model) உருவாக்கப்படுகிறது. ஒவ்வொரு பேச்சுக் குழல் துணை ஒழுங்கமைப்பின் ஒலியியக்க மாதிரியும் ஆயப்படுகிறது.

ஒரு நடைமுறை ஒலிப்பு கூட்டிணைப்பாக்கி (practical articulatory synthesizer) போச்சுக்குழல், எலும்பு உட்புழைகளுடன் கூடிய மூக்குக் குழல், குரல்வளைத் தடைகள், துணை குரல்வளை ஒழுங்குமுறை, கிளர்வுறல் மூலம், கிளர்தல் ஓசை மூலம் இவற்றை உட்படுத்தும். பேச்சு ஒழுங்குமுறையின் ஒலியியக்க சமன்பாடுகள் பரிந்துரைக்கப்பட்ட ஒலிப்பு கூட்டிணைப்பாக்கிக்கு வேண்டி உருவாக்கப்படுகின்றது. காலப்புல அணுகுமுறை (time-domain

approach) பேச்சு ஒழுங்குமுறையின் இயக்கப் பண்புக்கூறுகளைப் பாவிக்கப் பயன்படுத்தப்படுகின்றது. பேச்சுக் குழல் குறுக்குவெட்டுப் பரப்பு அல்லது ஒலிப்பு அளகைகள் (articulatory parameters), நேர்கோட்டு அல்லது ஆர்க் டான் செயல்பாட்டைப் (arc tan function) பயன்படுத்தி இரு அடுத்தடுத்த சட்டங்களுக்கு இடையில் செருகப்படுகின்றது. முறையாகத் திட்டமிடப்பட்ட ஒலிப்பு உருவாக்கிகள் உரசொலிகள் மற்றும் அடைப்பொலிகளின் எல்லா இயல்பான தேவையான விளைவுகளை மீள்ளுருவாக்கம் செய்யவும் உண்மையான பேச்சு உருவாக்கத்தில் நிகழும் பௌதிகச் செயல்பாடுகளை ஒக்கும் விதத்தில் இணை ஒலிப்பு மாற்றங்களையும் மூலக் குழாய் ஊடாட்டத்தையும் மாதிரிப்படுத்த இயலும்.

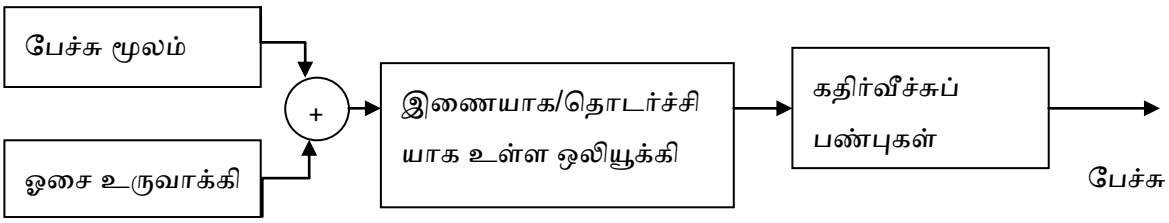
3.5.7.5.2 ஒலிச்செறிவுக் கூட்டிணைப்பாக்கம்

ஒலிச்செறிவு கூட்டிணைப்பாக்கம் (formant synthesis) ஓட்டநேரத்தில் மனிதப் பேச்சு மாதிரிகளைப் பயன்படுத்துவதில்லை. மாறாக ஒலியியக்க மாதிரிகளைப் பயன்படுத்தி வெளியீடு கூட்டிணைப்பாக்கம் செய்யப்பட்ட பேச்சு (output synthesized speech) உருவாக்கப்படுகின்றது. அடிப்படை நிகழ்வெண், குரல்படுத்தல் (voicing), இரைச்சல் நிலைகள் (noise levesl) போன்ற அளகைகள் செயற்கைப்பேசின் அலைவடிவை உருவாக்கும் படி கால அடிப்படையில் மாற்றப்படுகின்றது. இந்த அணுகுமுறை சிலவேளைகளில் விதி அடிப்படையிலான உருவாக்கம் (rule-based synthesis) என்று அழைக்கப்படுகின்றது; இருப்பினும் சிலர் பல ஒன்றிணைக்கும் ஒழுங்குமுறைகள் ஒழுங்குமுறையின் சில பகுதிகளுக்கு விதி அடிப்படையிலான உட்கூறுகளை/பகுதிகளை பயன்படுத்துகின்றன என்று வாதிடுகின்றனர்,

ஒலிச்செறிவு உருவாக்கத் தொழில்நுட்பம் அடிப்படையிலான பல ஒழுங்குமுறைகள் கீழ்வரும் படத்தில் காட்டியுள்ளது போல் செயற்கையான, ரபோட்டிக் ஓசைப் பேச்சு மற்றும் வெளியீட்டை உருவாக்குகின்றது; இருப்பினும் அதிக இயற்கைத்தன்மை பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறையின் (speech synthesis system) இலக்கு அல்ல; ஒலிச்செறிவு ஒழுங்குமுறைகள் (formant synthesis systems) ஒன்றிணைக்கும் ஒழுங்குமுறைகளைவிட (concatenative systems) சில அனுகூலங்களைக் கொண்டுள்ளன.

ஒலிச்செறிவு கூட்டிணைப்பாக்கம் செய்யப்பட்ட பேச்சு மிக அதிக வேகத்திலும் ஒன்றிணைப்பு ஒழுங்குமுறைகளைச் சிதைக்கும் ஒலியியக்கத் தடுமாற்றத்தை விலக்குவதால் நம்பகமான அளவில் புரியக்கூடியது. இரண்டாவது ஒலிச்செறிவுக் கூட்டிணைப்பாக்கிகள் ஒன்றிணைப்பு ஒழுங்குமுறைகளைவிட பெரும்பான்மையும் குறைந்த வழியமைப்புவரைவுகள் கொண்டது; ஏனென்றால் அவை பேச்சு மாதிரிகளின் தரவு அடிதளத்தைக் கொண்டிருக்கவில்லை.

நினைவக இடம், ஆய்வு ஆற்றல் பெரும்பான்மையும் அரிதாக இருக்கிற உட்படு கணிப்புச் சூழல்களிலும் (embedded computing situations) அவற்றைப் பயன்படுத்த இயலும். இறுதியாக ஒலிச்செறிவு அடிப்படை ஒழுங்குமுறைகள் வெளியீட்டுப் பேச்சின் எல்லா நோக்குகளுடனும் முழுக் கட்டுப்பாடு கொண்டிருப்பதன் காரணமாகக் கேள்விகளையும் கூற்றுகளையும் மட்டுமல்லாமல் பலவகை உணர்ச்சிகளையும் பேச்சின் சுரங்களையும் வெளிப்படுத்தும் மீக்கூறு அல்லது இசையோட்டத்தின் அதிக வகைகள் வெளியீடாக இருக்கவியலும்.



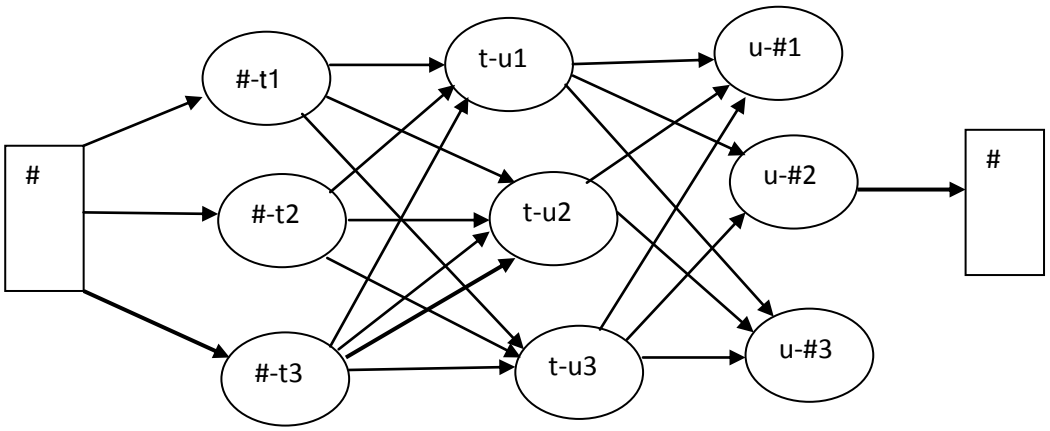
3.5.7.5.3 ஒன்றிணைப்பு கூட்டிணைப்பாக்கம்

ஒன்றிணைப்புப் பேச்சு கூட்டிணைப்பாக்கம் (concatenative synthesis) இயல்பான பேச்சிலிருந்து ஆக்கப்பட்ட வேறுபட்ட நீளமுள்ள ஏற்கனவே பதிவுசெய்யப்பட்ட பேச்சு அலகுகளைப் பயன்படுத்தும் செயல்பாடாகும். ஒன்றிணைப்பு உருவாக்கத்தில் முன்பதிவுசெய்யப்பட்ட பேச்சின் தரவுதளத்திலிருந்து உள்ள பேச்சுக் கூறுகள் ஒன்றிணைக்கப்படும். ஒன்றிணைப்பு பேச்சுக் கூட்டிணைப்பாக்க நுட்பங்கள் மூலம் பல வெற்றிகரமான ஒழுங்குமுறைகள் உருவாக்கப்பட்டுள்ளன. இருப்பினும் ஒன்றிணைப்பாக்கத்தில் பல சிக்கல் உள்ளன. அதிக அளபு இயற்கைத்தன்மையான ஒலிகளை உருவாக்கவேண்டி அவைகள் கவனத்தில் கொள்ளப்படவேண்டும். சில சிக்கல்கள் கீழே பட்டியலிடப்பட்டுள்ளன.

1. உருவாக்கப்பட்ட பேச்சில் இணைக்கப்படும் இடங்களில் தொடர்ச்சி இல்லாமை.
2. குறிப்பாக அசைகள், சொற்கள் போன்ற நீண்ட ஒன்றிணைப்பு அலகுகள் பயன்படுத்தப்படும்போது நினைவகத் தேவை பெரும்பான்மையும் அதிகமாகும்.
3. தரவைச் சேகரிப்பதற்கும் அடையாளப்படுத்துவதற்கும் பெரும்பான்மையும் காலம் அதிகமாகத் தேவைப்படும்.
4. வெளியீட்டுப் பேச்சு தேர்ந்தெடுக்கப்பட்ட தரவு அடித்தளத்தை அதிகமாகச் சார்ந்து அமையும்.

கோட்பாட்டு அடிப்படையில் எல்லா சாத்தியமான மாற்றொலிகளும் உட்படுத்தப்படவேண்டும்; இருப்பினும் தரத்திற்கும் மாதிரிகளின் எண்ணிக்கைகளுக்கும் இடையில் சலுகைகள் செய்துகொள்ளவேண்டும். ஒன்றிணைக்கப்படவேண்டிய பேச்சு அலகுகள் பதிவுசெய்யப்பட்ட ஒலியின் வேறுபட்ட இடங்களிலிருந்து பிரித்தெடுக்கப்படுவதால் ஒலியைப் பதிவு செய்யும் போது எப்போதும் இசைமை ஒன்றுபோல் இருக்க இயலாததாலும் வேறுபட்ட சொற்கள் வேறுபட்ட இசைமகளைக் கொண்டிருப்பதாலும் ஒன்றிணைக்கும் இடங்களில் தொடர்ச்சி இன்மை ஏற்படுகின்றது. இயல்பான ஒலியைப் பெறுவதற்காக இணைக்கப்படும் இடங்கள் அகற்றப்படவேண்டும்; இணைக்கப்படும் இரண்டு பேச்சு அலகுகளின் இசைமையைச் சமன்படுத்தி இதைச் செய்யவியலும்.

இந்தச் செயல்பாட்டில் நாம் தேவைப்படும் கூற்றுக்களை உருவாக்குவதற்காக முன்ஒலிப்பதிவு செய்யப்பட்ட தரவு அடித்தளத்திலிருந்து பேச்சின் கூறுகளைத் தெரிந்தெடுத்து ஒன்றோடு ஒன்று இணைக்கின்ற காரணத்தால் பின்வரும் படத்தில் காட்டப்பட்டுள்ள ஒன்றிணைப்பு உருவாக்கம் வெட்டி ஒட்டும் கூட்டிணைப்பாக்கம் (cut and paste synthesis) என்றும் அழைக்கப்படுகின்றது. இந்தச் செயல்பாட்டில் நாம் தேவைப்படும் கூற்றுக்களை உருவாக்குவதற்காக முன்ஒலிப்பதிவு செய்யப்பட்ட தரவு அடித்தளத்திலிருந்து பேச்சின் கூறுகளைத் தெரிந்தெடுத்து ஒன்றோடு ஒன்று இணைக்கின்றோம்.



ஒலிப்பதிவு செய்வதற்குப் பொதுவான தேர்வு ஒலியன்களும் ஒலியிருமைகளும் (diphones) ஆகும்; ஏனென்றால் அவைகள் போதுமான நெகிழ்வு பெறுவதற்கு வேண்டியும் நினைவகத் தேவைகளைப் போதுமான அளவுக்கு வைப்பதற்கு வேண்டியும் குறுகி உள்ளன. பல

காரணங்களால் அசைகள் அல்லது சொற்கள் போன்ற நீண்ட அலகுகளைப் பயன்படுத்துவது சாத்தியமில்லாதது மற்றும் நடைமுறைப் படுத்த இயலாதது ஆகும். ஒன்றிணைப்புகளில் ஒலியிருமைகளைப் (diaphones) பயன்படுத்துவது ஒலிப்புகளைக் கையாள நல்ல சாத்தியங்களைத் தருகின்றது; ஏனென்றால் ஒலியிருமைகள் ஒரு ஒலியனிலிருந்து மற்றொரு ஒலியனுக்கு மாறுவதையும் மற்றும் முதல் ஒலியனின் பின் பகுதியையும் இரண்டாம் ஒலியனின் முன் பகுதியையும் கொண்டிருக்கின்றது. இதன் விளைவாக ஒன்றிணைப்பு இடம் ஒவ்வொரு ஒலியனின் மையத்திலும் இடங்காணப்படும் மற்றும் இது ஒலியனின் நிலையான இடமாக இருப்பதன் காரணமாக எல்லைகளில் ஏற்படும் வேறுபாட்டின் அளவு குறைவதை எதிர்பார்க்க இயலும். ஒரு தரவு அடித்தளத்தின் வேறுபட்ட ஒலியன்களின் எண்ணிக்கை கிட்டத்தட்ட 40-இலிருந்து 50-க்குள் இருக்கையில் பொருத்தமுள்ள ஒலியிருமைகளின் எண்ணிக்கை இயல்பாக 1500-இலிருந்து 2000 வரைக்கும் இருக்கும்; இருப்பினும் இந்த அளவு தரவு அடித்தளத்தைக் கொண்ட கூட்டிணைப்பாக்கி பொதுவாகச் செயல்முறைப்படுத்தத் தக்கதாகும்.

பொதுவாக ஒருங்கிணைப்பாக்கம் மிக இயற்கையான உருவாக்கப்பட்ட பேச்சை உருவாக்கும். இருப்பினும் பேச்சில் உள்ள இயற்கையான வேறுபாடுகளுக்கும் அலைவடிவுகளைக் கூறிடுவதற்கான தானியக்கமான நுட்பங்களின் இயல்புக்கும் இடையிலான வேறுபாடுகள் வெளியீட்டில் கேட்கத்தக்க தடுமாற்றத்தில் விளையும். ஒன்றிணைப்பாகத்தின் மூன்று முக்கியத் துணை வகைகள் உள்ளன.

3.5.7.5.4 அலகுத் தேர்வு கூட்டிணைப்பாக்கம்

அலகுத் தேர்வு கூட்டிணைப்பாக்கம் (unit-selection synthesis) பதிவு செய்யப்பட்ட பேச்சின் பெரிய தரவு அடித்தளங்களைப் பயன்படுத்துகின்றது. தரவு அடித்தள உருவாக்கத்தின் போது ஒவ்வொரு பதிவுசெய்யப்பட்ட கூற்றும் பின்வருவனவற்றின் சிலவாக அல்லது எல்லாமாகக் கூறிடப்படும்: தனிப்பட்ட ஒலிகள், ஒலியிருமைகள், அரையொலிகள், அசைகள், உருபன்கள், தொடர்கள், மற்றும் வாக்கியங்கள். இயல்பாகக் கூறுகளாக்கும் பகுப்பு சிறப்பாக மாற்றப்பட்ட ஒரு பேச்சுப் புரிவான் குழுமத்தைப் பயன்படுத்தி அலைவடிவம், நிறமாலை நிழற்படம்/ஒலியியக்க வர்ணனைப்படம் போன்ற காட்சி உருப்படுத்தங்களைப் பயன்படுத்தி பின்னர் சில கையாலான திருத்தங்களுடன் ஒரு “கட்டாயப்படுத்தப்பட்ட வரிசையாக்க” முறைக்குச் செய்யப்படுகின்றது. பேச்சு தரவு அடித்தளத்தில் அலகுகளின் சொல்லடைவு, பகுப்பு மற்றும் அடிப்படை நிகழ்வெண் (இசைமை), காலம், அசையில் இருப்பிடம், அடுத்துவரும் ஒலிகள் போன்ற ஒலியியக்க அளகைகளின் அடிப்படையில் உருவாக்கப்படுகின்றது. வழிமுறைவரைவு

இயக்கக் காலகட்டத்தில் (runtime) விரும்பப்பட்ட இலக்குக் கூற்று தரவு அடித்தளத்திலிருந்து தேர்வை நாடும் அலகுகளின் நற் தொடர்ச்சியியை உறுதிசெய்வதால் உருவாக்கப்படுகின்றது. இந்தச் செயல்பாடு சிறப்பாக மதிப்பீடுசெய்யப்பட்ட தீர்மானக் கிளையமைப்பைப் பயன்படுத்தி வழக்கமாகச் சாதித்துக்காட்டப்படுகின்றது.

அலகுத் தேர்வு பதிவுசெய்யப்பட்ட பேச்சுக்கு இலக்கவெண்ணியல்/டிஜிட்டல் குறிகை/சமிக்கைப் பகுப்பாய்வின் (digital signal processing) குறைந்த அளவையே பயன்படுத்துவதன் காரணமாக அதிக இயற்கைத்தன்மையை வழங்குகின்றது. இலக்கவெண்ணியல் குறிகைப் பகுப்பாய்வு பெரும்பான்மையும் பதிவுசெய்யப்பட்ட பேச்சை குறைந்த அளவு இயல்பானதாகச் செய்கின்றது; சில ஒழுங்குமுறைகள் அலைவடிவை நேர்த்திசெய்ய ஒன்றிணைக்கும் இடத்தில் குறிகைப் பகுப்பாய்வின் குறைந்த அளவையே பயன்படுத்துகின்றன. நல் அலகுத்தேர்வு ஒழுங்குமுறைகளின் (best unit selection systems) வெளியீடு முக்கியமாகப் பனுவலிலிருந்து பேச்சு ஒழுங்குமுறை ஒத்திசைவு செய்யப்பட்ட சூழல்களில் பெரும்பான்மையும் உண்மையான மனிதக் குரல்களிலிருந்து வேறுபடுத்த இயலாதனவாய் உள்ளன. இருப்பினும் அதிகப்படியான இயற்கைத்தன்மை இயல்பாக அலகுத் தேர்வு பேச்சுத் தரவு அடித்தளத்தளங்களை மிகப் பெரிதாக வேண்டுகின்றது; சில ஒழுங்குமுறைகளில் பேசின் நீண்ட மணிநேரங்களை உருப்படுத்தம் செய்யும் கிகாபைட் அளவிலான பதிவு செய்யப்பட்ட தரவை வேண்டும். மேலும் தரவு அடித்தளத்தில் நற் தேர்வு சாத்தியம் இருந்தபோதிலும் குறைந்த பேச்சு கூட்டிணைப்பாக்கத்தை விளைவிக்கும் கூறுகளையே தேர்வுசெய்யும் வழக்கத்திற்கு அலகுத் தேர்வு வழிமுறை வரைவுகள் (unit selection algorithms) அறியப்படுவன ஆகும்.

3.5.7.5.5 ஒலியிருமை கூட்டிணைப்பாக்கம் (Diaphone synthesis)

ஒலியிருமைக் கூட்டிணைப்பாக்கம் (Diaphone synthesis) மொழியில் நேரும் எல்லா ஒலியிருமைகளையும் கொண்டிருக்கும் குறைந்த அளவிலான பேச்சுத் தரவு அடித்தளத்தைப் பயன்படுத்துகின்றது. ஒலியிருமைகளின் எண்ணிக்கை மொழியின் ஒலியன் வரன்முறையை (phonotactics) பொறுத்து அமையும். ஒலியிருமைக் கூட்டிணைப்பு உருவாக்கத்தில் ஒவ்வொரு ஒலியிருமைக்கும் ஒரேயொரு எடுத்துக்காட்டுதான் பேச்சுத் தரவு அடித்தளத்தில் இருக்கும். வழியமைப்பு இயக்கக் காலகட்டத்தில் ஒரு வாக்கியத்தின் இலக்கு மீக்கூறு இலக்கவெண்ணியல் குறிகை பகுப்பாய்வு நுட்பங்களின் மூலம் இந்தக் குறைந்த அலகுகள் மீது சுமத்தப்படும். விளைவிக்கப்படும் பேச்சு அலகுத்தேர்வு ஒழுங்குமுறைகளைக் காட்டிலும் பொதுவாக

மோசமானதாக இருக்கும்; ஆனால் ஒலிச்செறிவு கூட்டிணைப்பாக்கிகளின் வெளியீட்டைக் காட்டிலும் அதிக இயல்பானதாக இருக்கும். ஒலியிருமை கூட்டிணைப்பாகம் ஒன்றிணைப்பு கூட்டிணைப்பாகத்தின் ஒலிசார் தடுமாற்றங்களாலும் ஒலிச்செறிவு கூட்டிணைப்பாக்கத்தின் செயற்கை ஒலிப்பு இயல்பாலும் குறையுற்றிருக்கும்; சிறிய அளவு என்பதைத் தவிர இவ்விரு அணுகுமுறைகளிலிருந்தும் குறைந்த அளவு அனுகூலங்களையே கொண்டுள்ளது. எனவே இதன் வணிகவயமான பயன்பாடுகள் குறைகின்றன; இருப்பினும் பல மென்பொருள் நடைமுறைப்படுத்தல்கள் இலவசமாகக் கிடைப்பதன் காரணமாக இது தொடர்ந்து ஆய்வில் பயன்படுத்தப்பட்டு வருகின்றது.

3.5.7.5.6 பொருண்மைக்களச் சிறப்புக் கூட்டிணைப்பாக்கம் (Domain-specific synthesis)

பொருண்மைக்களச் சிறப்புக் கூட்டிணைப்பாக்கம் (Domain-specific synthesis) முழு கூற்றையும் உருவாக்க முன் ஒலிப்பதிவு செய்யப்பட்ட சொற்களையும் தொடர்களையும் ஒன்றிணைக்கின்றது. இது பயணக் கால அட்டவணை அறிவிப்புகள் அல்லது வானிலை அறிக்கைகள் போன்ற ஒரு குறிப்பிட்ட பொருண்மைக் களத்தை எல்லைப்படுத்தும் வெளியீடுகளின் பயன்பாடுகளில் பயன்படுத்தப்படுகின்றது. இத்தொழில் நுட்பம் நடைமுறை படுத்துவதற்கு எளியது மற்றும் பேசும் கடிகாரங்கள் கணிப்பான்கள் போன்ற கருவிகளில் நீண்டகால வணிகப் பயன்பாடு கொண்டுள்ளது. வாக்கிய வகைகள் எல்லைக்குட்பட்டு இருப்பதாலும் மற்றும் அவை மூல ஒலிப்பதிவுகளின் மீக்கூறு மற்றும் இசையோட்டத்துடன் நெருக்கமாகப் பொருந்துவதாலும் இவ்வொழுங்கமைப்புகளின் இயற்கைத்தன்மை நிலை மிகவும் உயர்ந்ததாய் இருக்க இயலும்.

இவ்வொழுங்கமைப்புகள் அவற்றின் தரவு அடித்தளங்களில் சொற்கள் மற்றும் தொடர்களால் எல்லைப்படுத்தப்பட்டுள்ளதாலும் அவை பொது நோக்கத்திற்கு உகந்ததல்ல மற்றும் அவை முன் வழியமைப்பு செய்யப்பட்ட சொற்கள் மற்றும் தொடர்களின் ஒன்றிணைப்புகளை மட்டுமே கூட்டிணைப்பு உருவாக்கம் செய்ய இயலும்.

3.5.7.5.7 எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பு உருவாக்கம் (HMM based synthesis)

எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பாக்க அணுகுமுறை (HMM based synthesis) ஹிட்டன் மார்க்கோவ் மாதிரிகள் அடிப்படையிலானது. இந்த ஒழுங்குமுறையில் நிகழ்வெண் நிறமாலை/அலைமாலை (பேச்சுக் குழாய்), அடிப்படை நிகழ்வெண் (பேசு மூலம்), பேச்சின் கால அளவு (மீக்கூறு) என்பன எச்.எம்.எம்.களால் ஒரே நேரத்தில் மாதிரிப்படுத்தப்பட்டுள்ளன. பேச்சு

அலைகள் கூடுதல் சாத்திய அளவீடு (maximum likelihood criterion) அடிப்படையில் உருவாக்கப்படுகின்றது.

இவ்வொழுங்கமைப்பு இரண்டு நிலைகளைக் கொண்டுள்ளது: பயிற்சி நிலை, (training stage) கூட்டிணப்பாக்கம் (சோதனை) (testing). பயிற்சி நிலையில் ஒலியன் எச்.எம்.எம்.கள் பேச்சு தரவு அடித்தளத்தைப் பயன்படுத்திப் பயிற்சி செய்யப்படுகின்றது. நிறமாலை/அலைமாலை (spectrum), அடிப்படை நிகழ்வெண் (F0) என்பன பல்லொழுக்கு எச்.எம்.எம்.களால் மாதிரிப்படுத்தப்பட்டுள்ளன; இதில் நிறமாலை/அலைமாலை மற்றும் F0 பகுதிகளுக்கு வெளியீட்டு வினியோகங்கள் (output distributions) முறையே தொடர்ச்சியான நிகழ்வுத்தகவு வினியோகம் (continuous probability distribution) மற்றும் பல் இடைவெளி நிகழ்வுத்தகவு விநியோகம் (multi-space probability distribution) இவற்றைப் பயன்படுத்தி மாதிரிப்படுத்தப்பட்டுள்ளன

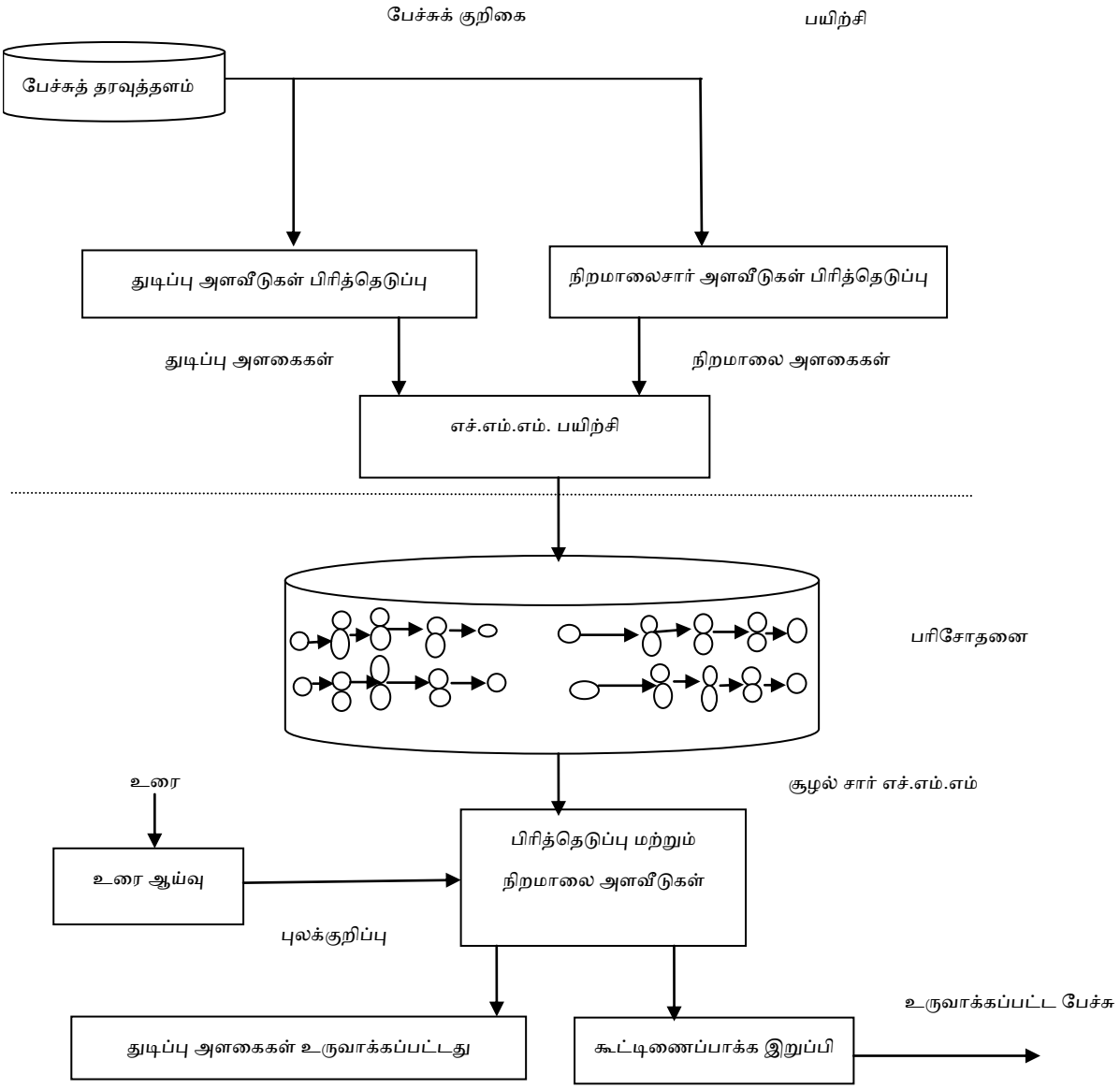
நிறமாலை/அலைமாலை மற்றும் F0 இவற்றின் வேறுபாடுகளை மாதிரிப்படுத்த ஒலியியல், மீக்கூறு, ஒலியன் அடையாளக் காரணிகள், அழுத்தம் தொடர்பான காரணிகள், இடம்சார் காரணிகள் போன்ற மொழிச் சூழல் காரணிகள் கருத்தில் கொள்ளப்படுகின்றன. பின்னர் தீர்மானக் கிளை அடிப்படையிலான சூழல் கொத்தாக்க நுட்பம் (decision tree based context clustering technique) சூழல் சார் ஒலியன் எச்.எச்.எம்.களின் நிறமாலை/அலைமாலை மற்றும் F0 பகுதிகளுக்குத் தனித்தனியாகப் பயன்படுத்தப்படுகின்றது. இறுதியாக நிலை கால அளவுகள் பல்பரிமாண காசியன் வினியோகங்களால் (multi-dimensional Gaussian distributions) மாதிரிப்படுத்தப்படுகின்றது மற்றும் நிலைக் கொத்தாக்க நுட்பம் (state clustering technique) கால அளவு மாதிரிகளுக்குப் பயன்படுத்தப்படுகின்றது. எச்.எம்.எம்.இன் கட்டு வரைபடம் கீழே தரப்பட்டுள்ளது.

கூட்டிணைப்பாக்க (பரிசோதனை) நிலையில் தரப்பட்ட ஒரு பனுவல் சூழல் சார் ஒலியன் புலக்குறிப்பு தொடர்ச்சியாக மாற்றப்படுகின்றது. புலக்குறிப்புத் தொடர்ச்சி அடிப்படையில் ஒரு வக்கியம் எச்.எம்.எம். சூழல்சார் ஒலியன் எச்.எச்.எம்.களை ஒன்றிணைப்பதால் உருவாக்கப்படுகின்றது. ஒலியன் கால அளவுகள் நிலை கால அளவு வினியோகங்களைப் பயன்படுத்திப் பெறப்படுகின்றது; பின்னர் நிறமாலை/அலைமாலை மற்றும் F0 அளகை தொடர்ச்சிகள் வாக்கிய எச்.எம்.எம்.-இலிருந்து எம்.எல். அளவீடு அடிப்படையில் பெறப்படுகின்றது. இறுதியாக எம்.எல்.எஸ்.எ.-ஐப் பயன்படுத்தி உருவாக்கப்பட்ட மெல்-

செப்ட்ரல் (Mel-septral) மற்றும் F0 அளகை தொடர்ச்சிகள் இவற்றிலிருந்து இறுப்புப் பேச்சு (filter speech) கூட்டிணைப்பாக்கப்படுகின்றது.

இந்த அணுகுமுறையில் பேசும் பாணிகள்/நடைகள் இரண்டு வழிகளில் மாதிரிப்படுத்தப்பட்டுள்ளது: "நடை/பாணி சார் மாதிரியாக்கம்" அல்லது "நடை கலவை மாதிரியாக்கம்". நடைசார் மாதிரியாக்கத்தில் ஒவ்வொரு நடையும் தனித்தனியாக மாதிரிப்படுத்தப்படுகின்றன. மாறாக நடைக் கலவை மாதிரியாக்கத்தில் பேசும் பாணிகளும்/நடைகளும் உணர்ச்சி வெளியீடுகளும் சூழல் காரணிகளாகவும் ஒலியியல் மற்றும் மொழியியல் காரணிகளாக எடுத்துக்கொள்ளப்படுகின்றது; மற்றும் எல்லா நடைகளும் ஒரே நேரத்தில் ஒரு தனியான ஒலியியக்க மாதிரியால் மாதிரிப்படுத்தப்பட்டுள்ளன. இவ்விரு மாதிரியாக்க அணுகுமுறைகளுக்கும் ஒரே நிறைவேற்றத்தைக் கொண்டிருக்கும் மற்றும் பதிவு செய்யப்பட்ட பேச்சுடன் ஒற்றுமையுள்ள பேச்சைச் கூட்டிணைப்பாக்கம் செய்வது சாத்தியமாகும்.

நம்மிடம் வேறுபட்ட இரு நபர்களின் இரு பேச்சு நடை மாதிரிகள் இருந்தால் இரு நபர்களின் மாதிரிகளை இடைச் செருகுவதால் இரு பேசுபவர்களுக்கு இடையில் இடைப்பட்ட குரல் சிறப்புப்பண்புகளைக் கொண்ட பேச்சைக் கூட்டிணைப்பாக்கம் செய்ய இயலும். இதற்கு நாம் கவனக்கணிப்புகளுக்கு (observation) இடைச்சொருகல் எனப்படும் எளிய இடைச்சொருகல் நெறிமுறையைப் (interpolation method) பயன்படுத்துகின்றோம்.



3.5.7.5.8. பேச்சு கூட்டிணைப்பாக்க நெறிமுறைகளின் நிறைகளும் குறைகளும்

கூட்டிணைப்பாக்க வகை	***நிறை***	**குறை**
ஒலிப்பு கூட்டிணைப்பாக்கம்	பேச்சின் பௌதிகப் பாவனை; ரபோட்டிக் பயன்பாடுகளுக்குப் பொருத்தமானது.	இயல்புத்தன்மை/ இயற்கைத் தன்மை குறைவு

ஒலிச்செறிவு கூட்டிணைப்பாக்கம்	வேறுபட்ட பேச்சு நடைகளுகளை உட்படுத்தும் நெகிழ்வுத்தன்மை.	அளகைகளின் கையிலான இசைவிப்பு
ஒன்றிணைப்பு கூட்டிணைப்பாக்கம்	நல்ல இயற்கைத் தன்மை.	குறைவான புரிந்துகொள்கை; அதிக நினைவகம் தேவைப் படுதல்.
எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பாக்கம்	நல்ல பண்புள்ள புரியவியலும் பேச்சைப் பெற குறைந்த அளவிலான நினைவகத் தேவைகள்	அலகுத் தேர்வுடன் ஒப்பிடுகையில் இயற்கைத் தன்மையின் தரக்குறைவு.

3.5.7.5.6. உரை-பேச்சு ஒழுங்குமுறைகளின் பயன்பாடுகள்

உரையிலிருந்து பேச்சு ஒழுங்குமுறை இயல்பான மொழியை (உரையை) பேச்சாக மாற்றுகின்றது. தற்காலத்தில் உரையிலிருந்து பேச்சு ஒழுங்குமுறை பல நோக்கங்களுக்காகப் பயன்படுத்தப்படுகின்றது. உரையிலிருந்து பேச்சு ஒழுங்குமுறை சில பயன்பாடுகளுக்குப் பயன்படுத்தப்படுகின்றது. இதன் சில பயன்பாடுகள் கீழே பட்டியலிடப்பட்டுள்ளன.

1.இலக்கவெண்சார் நூல் படிப்பான் (digital reader)

2.கைபேசி செய்தி படிப்பான் (mobile news reader)

3.பேச்சு ஒழுங்குமுறை (speech system)

4. கேளிக்கை/பொழுதுபோக்குத் தொழில் (entertainment industry)

3.5.7.5.6.1. இலக்கவெண்சார் நூல்படிப்பான்

கணிப்பொறி திரையில் உரைகளைப் படிப்பதற்கு நிறைய காலம் எடுத்துகொள்கின்றது. பயன்படுத்துபவர் மின் அஞ்சலைப் படிப்பர், ஸ்ப்ரெட்ஷீட்களை நிரப்புவர், இணைய தளத்திலிருந்து தகவல்களைச் சேகரிப்பர், சொல்லாய்வியில் கட்டுரைகளைத் திருத்துவர்; இது போன்று பல செயல்களைச் செய்வர். பெரும்பாலும் எல்லோரும் கண் களைபுக்கு ஆளாயிருப்பர். உரையிலிருந்து பேச்சு ஒழுங்குமுறை இச்செயல்களை எளிதாக்குகின்றது. அது கணியில் உள்ள உரைகளை கேட்கக்கூடிய வடிவில் மாற்றித் தருகின்றது.

3.5.7.5.6.2 கைபேசி செய்தி படிப்பான்

எல்லோருக்கும் பல வேலைகள் இருக்கும். அவர்கள் தங்கள் இயல்பான வேலைகளில் ஈடுபடுவதற்கு இடையில் செய்திகளை அறிய விரும்புவர். இன்றைய கால கட்டத்தில் எல்லோர்

கையிலும் கை பேசி உள்ளது. உரையிலிருந்து பேச்சு ஒழுங்குமுறையைக் கைபேசியில் நிறுவினால் செய்திகளை உரை-பேச்சு ஒழுக்குமுறை படிக்கக் கேட்கலாம்.

3.5.7.5.6.3. பேச்சு ஒழுங்குமுறை

உரை-பேச்சு ஒழுங்குமுறை பேச்சு ஒழுங்குமுறையின் பகுதியாகும். இது ஒரு மொழியில் எழுதப்பட்டுள்ள உரைகளை இலக்கு மொழியாக மாற்றுகின்றது. இறுதியில் பேச்சு-உரை ஒழுங்குமுறை எழுத்து உரையை பேச்சு உரையாக மாற்றித் தரும்.

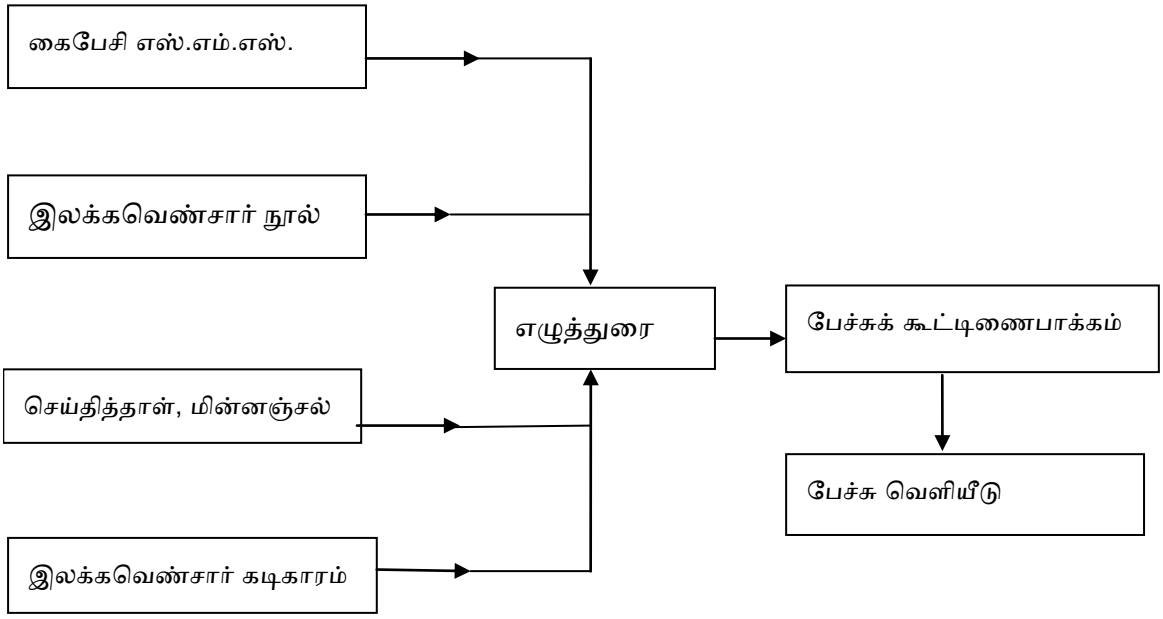
3.5.7.5.6.4 மருத்துவப் புலம்

உரை-பேச்சு ஒழுங்குமுறை தொழிநுட்பத்திற்கு உதவும் கருவிகளாகப் பயன்படுகின்றது. மிகப் பெரிய பயன்பாடு பார்வை குறைபாடு உள்ளவர்களுக்குத் திரை படிப்பானாகப் பயன்படுவது. உரை-பேச்சு ஒழுங்குமுறை பார்வையற்றவர்களுக்கு நூல்களைப் படித்துக் காட்டும்; மின்னஞ்சல்களைப் படித்துக்காட்டும்; இணைய தளத்திலிருந்து வேண்டிய செய்திகளையும் தகவல்களையும் படித்துக் காட்டும். மட்டுமன்றி படிகவியலாத அல்லது படிக்க முயற்சிக்கும் குழந்தைகளுக்கு படித்துக் காட்டும் பயிற்சியாளராகவும் பயன்படுகின்றது. பேச்சுக் குறைபாடு உள்ளவர்களுக்கும் பயிற்சியாளராகப் பயன்படுத்த இயலும்.

3.5.7.5.6.5. கேளிக்கைத் தொழில்

இன்றைய கால கட்டத்தில் உரை-பேச்சு ஒழுங்குமுறை பல்லூடக தொழிலுக்குப் பெரிதும் பயன்படுகின்றது. உரை-பேச்சு ஒழுங்குமுறை கணினி விளையாட்டு, வீடியோ விளையாட்டு மற்றும் பிற பல கேளிக்கைத் தொழில்களில் பயன்படுகின்றது. ஒட்டுமொத்த பயன்பாட்டு மாதிரி கீழே காட்டப்பட்டுள்ளது.

படம் 10



3.5.8. தமிழில் உரையிலிருந்து பேச்சாக்கம்

பேச்சு மற்றும் கட்புலன் போன்ற மனித வசதிகளை இயந்திரத்தில் உட்படுத்துவது செயற்கை அறிவு ஆய்வின் அடிப்படை விசயமாகும். பேச்சு வெளியீட்டை உருவாக்கும் கணிப்பொறியின் திறன் பேச்சுக் கூட்டிணைப்பு உருவாக்கம் என அழைக்கப்படுகின்றது. இது பேச்சு உற்பத்தி மற்றும் புலனுணர்வின் ஆழமான புரிந்துகொள்ளலை வேண்டுகின்றது. கணிப்பொறிக்கு மனித இடைமுகத்தின் முன்னேற்றத்தைப்பற்றிப் பரவலான பேச்சு இருக்கின்றது. மக்கள் தேவையான தரவை உட்கார்ந்து தட்டச்சு செய்ய விரும்பவில்லை அல்லது கணினித் திரையிலிருந்து தரவைப் படிக்க விரும்பவில்லை. இந்நோக்கத்தில் பேச்சு கூட்டிணைப்பு உருவாக்கம் கணிப்பொறிக்கு மனித இடைமுகத்தை மேம்படுத்த முக்கியமான படிகளில் ஒன்றாக அமைகின்றது. இவ்வியலில் தமிழுக்கு உரையிலிருந்து/பனுவலிலிருந்து பேச்சு கூட்டிணைப்பாக்கத்தை (speech synthesis) நடைமுறைப்படுத்துவது பற்றி கூறப்படும்.

3.5.8.1. தமிழில் உரையிலிருந்து பேச்சாக்கத்தின் பின்னடைவு

தமிழில் உரையிலிருந்து/பனுவலிலிருந்து பேச்சை உருவாக்கும் முயற்சி பின்வரும் காரணங்களால் பின்னடைவு காட்டுகின்றது.

1. தமிழ் மொழியின் கலவைத் தன்மை
2. தமிழ் மொழி இலக்கணத்தின் கலவைத் தன்மை (ஒலியன் வரன்முறை, மெய்மயக்கங்களின் அமைப்பு, புணர்ச்சி மாற்றங்கள்)

3. தூயதமிழ்ச் சிக்கல் (பேச்சுத் தமிழுக்கும் தூய எழுத்துத் தமிழுக்கும் உள்ள வேறுபாடுகள்)
4. ஏற்கனவே உருவாக்கப்பட்ட தமிழ் உரையிலிருந்து/பனுவலிலிருந்து பேச்சு ஒழுங்கமைப்புகளின் செயற்கைத் தன்மை.

3.5.8.2. தமிழில் உரையிலிருந்து பேச்சாக்கத்தின் முயற்சிகள்

தமிழில் உரையிலிருந்து பேச்சாக்கத்தின் முயற்சிகளில் முதன்மையாக குறிப்பிடத்தகுந்தவர்கள் யக்ஞநாராயணாவும் (Yegnanarayana et al., 1994) ஹேமா மூர்த்தியும் ஆவர். தமிழ் மொழிகளுக்கான பேச்சுக் கூட்டிணைப்பாக்கத்திற்காக ஆய்வுகள் மேற்கொண்டு கட்டுரைகளாக வெளியிட்டுள்ளனர். ஹேமா மூர்த்தி பல ஆய்வுத்திட்டங்களை மேற்கொண்டு தமிழில் உரையிலிருந்து பேச்சாக்கத்திற்கான ஒழுங்குமுறையை உருவாக்கியுள்ளார்.

தமிழ் உரையிலிருந்து பேச்சாய்வில் இந்திய அறிவியல் நிறுவனத்தைச் சர்ந்த ராமகிருஷ்ணாவின் பங்களிப்பு இங்கு கட்டாயம் குறிப்பிட வேண்டிய ஒன்றாகும். உரையிலிருந்து பேச்சு உருவாக்கத்தில் அவர் பெரும் பங்களிப்பு செய்துள்ளார். பிரதிபாவும் அவரும் எழுதிய "தமிழுக்கான இணையச் சாத்தியமான பேச்சு உருவாக்கி (Prathibha and Ramakrishana 2002), அவரும் பிறரும் (Ramakrishnan, Kaushik, Laxmi Narayana 2007) எழுதிய "தமிழ் உரையிலிருந்து பேச்சுக்கு இயற்கைமொழி ஆய்வு" மற்றும் அவர் எழுதிய பேச்சு நுட்பமும் தமிழும் (Ramakrishana 2014) என்பன சான்றுகளாகும்.

ஜெயவர்த்தனா மற்றும் பிறர் (Jayavardhana Rama et al 2001) திருக்குறளுக்கு ஒரு உரையிலிருந்து பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறை உருவாக்கியுள்ளனர். அவர்கள் தமிழ் மொழியில் உரையிலிருந்து பேச்சு மாற்றத்திற்கு ஒரு புதிய நெறிமுறையை விளக்குகின்றனர். இது இரண்டு நிலைகளை உள்ளடக்கியது: நேரடித்தொடர்பில்லா நிலை/ஆப்லைன் நிலை (offline phase) மற்றும் நேரடித்தொடர்பு நிலை/ஆன்லைன் நிலை (online phase). நேரடித்தொடர்பில்லா நிலை முன்செயற்பாங்கு (preprocessing), கூறுபடுத்தல் (segmentation), இசைமை அடையாளப்படுத்தல் (patch marking) இவற்றை உட்படுத்தும். மீக்கூறு (prosody) கூற்று வெளிப்பாட்டிற்குப் பயன்படுத்தப்படும் பௌதீக, ஒலியியல் விளைவுகளின் கலவைத்தன்மையான அலையாகும். மீக்கூறு இசைமை (pitch), விட்டிசை (pause), மாத்திரை (duration), ஒலிவன்மை/ஒலிமிகைப்பு (loudness) இவற்றை உள்ளடக்கும். இசைமை பேச்சு குறிகையின் அதிக வெளிப்படுத்தும் பகுதி ஆகும். கூட்டிணைப்பாக்கம் செய்யப்பட்ட பேச்சின் இயல்புத்தன்மையைப் பாதிக்கும் முக்கியக் காரணி மாத்திரை ஆகும். ஒரு சொலின் அல்லது வாக்கியத்தின் வேறுபட்ட இடங்களில் வரும் ஒரே உயிரொலி வேறுபட்ட மாத்திரைகளைக்

கொண்டிருக்கும். இதற்குப் பின்வரும் எடுத்துக்காட்டு தரப்பட்டுள்ளது: *நான் ஆறு மணிக்கு வரலாமா*. இந்த வாக்கியத்தில் உயிரொலி /ஆ/ வேறுபட்ட இடங்களில் வேறுபட்ட மாத்திரையில் வருகின்றன. இசைமை அடையாளத்தில் அலைவடிவங்கள் ஒன்றிணைக்கப்படுவதால் இசைமையை அடையாளப்படுத்தல் முக்கியமானதாகும். இசைமையை மதிப்பீடு செய்யத் தானியக்கப் பொருத்த நெறிமுறை (auto-correlation method) பயன்படுத்தப்பட்டுள்ளது. கால அளவின் நீட்சி தந்துள்ள கூறில் அதிகபட்ச தானியக்கப் பொருத்தத்தைக் கண்டுபிடிப்பதால் கணிக்கப்படுகின்றது. உரை/பனுவல் ஆய்வு நிலை பேச்சின் அடிப்படை அலகுகளின் கோர்வைகளாகப் பகுப்பாய்வு செய்வதையும் தமிழ் விதிகளைப் பயன்படுத்துவதையும் உட்படுத்தும். கூட்டிணைப்பாக்க நிலை சரியான வரிசைகிரமத்தில் இவ்வலகுகளின் அலைவடிவுகளின் ஒன்றிணைப்பை உட்படுத்தும். இயற்கைத் தன்மை அடிப்படையில் இத்திட்டம் விரிவுபடுத்தபட்டுள்ளது.

பேச்சுக் கூட்டிணைப்பாக்கம் மற்றும் அவற்றின் உயர்ந்த நிலைப் பயன்பாடு தொடர்பாகத் தமிழ் உரை-பேச்சு இயந்திரத்தின் நடைமுறையாக்கம், முன்னேற்றம், நடைமுறைப்படுத்தலின் சவால்கள் இவற்றின் பல்வேறு நெறிமுறைகள் சுருக்கமாகக் கூறப்பட்டுள்ளது. பனுவலிலிருந்து-பேச்சு ஒழுங்குமுறையின் நடைமுறையாக்கம் ஒன்றிணைப்பு நெறிமுறையைப் பயன்படுத்தி செய்யப்படுகின்றது; பனுவலிலிருந்து பேச்சு இயந்திரத்தின் தேவையான பகுதிகளான சொல் பகுப்பாய்வி (lexical analyzer), ஒலியன் கண்டுபிடிப்பி (phoneme identifier), குரல் பொருத்துதல் (voice mapping), பேச்சு கூட்டிணைப்பாக்கி (speech synthesizer), குரல் சீராக்கம் (voice modulation) இவைபற்றி விளக்கப்பட்டுள்ளது. இங்குப் பயன்படுத்தப்படும் சொல்லாய்வி வழக்கமான சொல்லாய்வியுடன் ஒற்றுமை காட்டுகின்றது; அவை தமிழ் எழுத்துருக்களை உட்படுத்த இயலும் (எ.கா. முரசு அஞ்சல்); பனுவலை/உரையை ஒலியியல் அடிப்படையில் உள்ளீடு செய்யும் அல்லது தமிழ் விசைப்பலகையைப் பயன்படுத்தும் விருப்பத்தேர்வு கொண்டிருக்கின்றது. இது தரவு பேச்சாக வெளிப்படுத்தப்பட வேண்டும் என்று விரும்புகிற பயன்படுத்துபவரிடமிருந்து உரை/பனுவல் பதிவைப் பெறும். ஒவ்வொரு சொல்லுக்கும் பின் வேகம், கன அளவு போன்ற மதிப்புகள் வழுநிலையாக வைக்கப்படும்.

எ.கா.

/s[75]/v[75]அ/tன்னைs/[75]ஒ/tரு/s[75]/v[75]தெ/tய்வம்

இது சொற்களை 75% வேகத்திலும் 75% கன அளவிலும் பொருத்தமான அசை அழுத்தத்திலும் இசைக்கும். சொல்சார் அசை அழுத்த ஆய்வி சுவையற்ற விளைவை நீக்க உதவ வேண்டி

வாக்கியத்தின் பல்வேறு பாகங்களுக்கு வெளிப்படையாகக் கூறப்படாதவை தவிர பிறவற்றிற்குச் சரியான அசை அழுத்தத்தைச் சேர்ப்பதற்காகத் தந்துள்ள பனுவலை தமிழ் இலக்கணத்திற்காகப் பரிசோதிக்கும். ஒலியன் அடையாளம் கண்டுபிடிப்பியால் *அன்னை* என்ற சொல் Aது NதுNAI என்று பகுப்பாய்வு செய்யப்படும்; இதில் ஒவ்வொன்றும் அவற்றின் தொடர்புள்ள ஒலிகளைக் கொண்டிருக்கும். இதுபோன்று எல்லா சொல்லையும் பகுத்தாயலாம்.

கிஷோர் மற்றும் பிறர் (Kishore, Rohit Kumar and Rajeev Sangal 2002) ஒன்றிணைப்புக்கு அசைகளை அடிப்படையாகப் பயன்படுத்தி தரவு-இயக்க அணுகுமுறை/நெறிமுறை (data-drivan method) அடிப்படையில் இந்திய மொழிகளுக்கு உரையிலிருந்து பேச்சுக் கூட்டிணைப்பாக்கத்திற்கு முயன்றுள்ளனர். இந்நெறிமுறையின் அலகுத்தேர்வு வழிமுறைவரைவு மீக்கூறு பொருத்தச் செயற்பாட்டைக் கையாளும் அனுகூலத்தைக் கொண்டுள்ளது; இது சொற்கள், தொடர்கள், வாக்கியங்கள் போன்ற பெரிய வரிசைகளை உட்படையாகத் தேர்ந்தெடுக்க இயலும். முன்மொழியப்பட்ட இந்த அணுகுமுறை இந்திய மொழிகளுக்கு உயர்ந்த சிறப்புப்பண்புள்ள பேச்சைத் தருவதாகக் கூறுகின்றனர்.

கிருதிகா, கீதா ஆகியோர் (Krithiga M.V. and Geetha T.V. 2004) எல்.பி.சி. அடிப்படையிலான ஒலியிருமை தரவு அடித்தளம் (diaphone database) மற்றும் கால அளவை மாற்றம் வழி அசை இசைமையின் மாற்றம் இவற்றைப் பயன்படுத்தி தமிழ் பனுவலிலிருந்து பேச்சு உருவாக்கத்தின் பண்பில் முன்னெற்றத்தைப் பற்றி விளக்குகின்றனர். ஒன்றிணைக்கும் பேச்சு கூட்டிணைப்பாக்கியால் (concatenative speech synthesizer) பேச்சு உருவாக்கப்படுகின்றது. அசை அலகுகள் அவற்றின் பண்பைக் குறைக்காமல் அலகு எல்லைகளில் ஸ்பெக்டரல் தொடர்ச்சியின்மையைக் குறைகும் படி ஒன்றிணைக்கப்படவேண்டும். ஒன்றிணைப்பு எல்லையில் பொருத்தமான ஈரொலியை செருகுவதாலும் கால அளவை மற்றத்தை நிறைவேற்றுவதாலும் நேர்த்தி ஏற்படுத்தப்படுகின்றது.

இந்திய மொழிகளின் தொழிநுட்ப முன்னெற்றம் (Technological Development of Indian Languages) என்ற இந்திய அரசின் ஆய்வித்திட்டத்தின் கீழ் தமிழ் மொழிக்காக 'எதிரொலி' என்ற ஒழுங்குமுறை சென்னை அண்ணாப் பல்கலைக்கழத்தால் உருவாக்கப்பட்டுள்ளது.

சாமுவல் தாமஸ் மற்றும் பிறர் (Samuvel Thomas, Nageshwara Rao, Hema A. Murthy, Ramalingam) 'Natural Sounding Tts Based On Syllable-Like Units' என்ற தமது கட்டுரையில் ஒன்றிணைப்பாக்கப் பேச்சுக் கூட்டிணைப்பாக்கத்திற்குப் புதிய அசைபோன்ற அலகைப் பயன்படுத்துவது பற்றி விளக்குகின்றனர். இந்த அலகுகள் குழும தாமத அடிப்படையிலான

பிரிப்பு வழிமுறைவரைவு (delay based segmentation algorithm) மற்றும் ஒலியியக்க அடிப்படையில் C*VC* (C: consonant, V: vowel) வடிவுக்குப் பொருந்த்துதல் இவற்றைப் பயன்படுத்தி தனியக்கமாக உருவாக்கப்படுகின்றன. இந்த ஒழுங்குமுறையின் திறன் தமிழில் இயல்பாக ஒலிக்கும் பேச்சைக் கூட்டிணைப்பாக்கம் செய்து நிரூபிக்கப்பட்டுள்ளது. ஓரசை அலகுகளுக்குப் பதிலாக ஈரசை அலகுகளைப் பயன்படுத்தியபோது கூறத்தக்க தர முன்னேற்றம் கிடைக்கப்பெற்றது; இதன் விளைவு மரபு ஒலியிருமை அடிப்படையிலான அணுகுமுறையைவிடச் மிகச் சிறந்ததாக அமைந்தது. இந்த அணுகுமுறையின் முக்கியமான சாதகநிலை மீக்கூறு விதிகளை நீக்கிவிடுவதாகும். பெறப்பட்ட பேச்சு இயல்பாக அமைவது இந்த அணுகுமுறையின் சிறப்பாகும். (என்பன இவர்களின் வாதம்).

வேணுககோபாலகிருஷ்ணா மற்றும் பிறர் (Venugopalakrishna Y R, மற்றும் பிறர் 2008) ‘Design and Development of a Text-To-Speech Synthesizer for Indian Languages’ என்ற தமது கட்டுரையில் ஒன்றிணைப்பாக்கத்தின் அலகுகளாக அசைகளையும் பல்லசைகளையும் கொண்டு அலகுத் தேர்வு அடைப்படையிலான உரை-பேச்சு கூட்டிணைப்பாக்கியின் திட்டவரைவு மற்றும் நடைமுறைப் படுத்தல் பற்றி விளக்குகின்றனர். இந்திய மொழிகள் அசையை மையமாகக் கொண்டவை ஆதலால் இந்திய மொழிகளுக்கு அசையை அலகாகத் தெரிவு செய்தது பொருத்தமானதாகும். அசை அடிப்படையிலான கூட்டிணைப்பாக்கம் கூறத்தக்க மீக்கூறு மாற்றத்தை வேண்டாவிட்டாலும் அசையின் சூழலில் நிறைவேற்றப்படவேண்டிய மீக்கூறு மாற்றம் மரபு ஒலியிருமை அடிப்படையிலான கூட்டிணைப்பாக்கதைவிட கூறத்தக்க விதத்தில் வேறுபட்டதாகும். (என்பன இவர்களின் வாதம்).

வினோத் மற்றும் பிறர் (Vinodh M.V., Bellur A., Narayan, B.K., Thakare D. M., Susan A., Suthakar N.M. and Murthy H.A.) ‘Using polysyllabic units for text to speech synthesis in Indian languages’ என்ற தமது கட்டுரையில் பல்லசை அலகுகளைப் (polysyllabic units) பயன்படுத்தி இந்திய மொழிகளுக்கு உரை-பேச்சு ஒழுங்குமுறையை திட்டமிடுவது மற்றும் உருவாக்குவது குறித்து விளக்குகின்றனர். முதலில் ஒலியன் அடிப்படையிலான உரை-பேச்சு ஒழுங்குமுறை உருவாக்கப்பட்டது. பின்னர் உரை-பேச்சுக்கு ஓரசை கொத்து அலகு (monosyllable cluster unit) உருவாக்கப்பட்டது. (பொருத்தமான அலகுகள் கிடைத்தால்) பல்லசை அலகுகள் பயன்படுத்தப்படும் போது கூட்டிணைப்பாக்க வாக்கியத்தின் தரம் அதிகரிக்க இயலும் என்பது கண்டுணரப்பட்டது; ஏனென்றால் இம்மாதிரியான நேர்வுகளில் சாரொலிப்பின் (co-articulation) விளைவு தக்கவைக்கப்படும். எனவே இந்தி மற்றும் தமிழுக்கு ஒன்றுக்கும் மேற்பட்ட வகை மயக்க

அலகுகள் கொண்ட பல்லசை அலகுகள் உருவாக்கப்பட்டன. ஒழுங்குமுறை அலகுத் தேர்வு செயற்பாங்கின் போது அலகுகளில் சிறந்ததைத் தெரிந்தெடுக்கும்; இதன்காரணமாகச் சேர்க்கும் போதும் ஒன்றிணைக்கும் போதும் விளைகின்ற இழப்பைக் குறைக்க இயலும். ஆரம்பக்கட்ட கேட்டல் பரிசோதனை பல்லசை உரை-பேச்சு ஒழுங்குமுறை சிறந்த தரம் உள்ளது என்பதைச் சுட்டிக்காட்டுகின்றது. (என்பன இவர்களது வாதம்).

வாசுரங்கநாதனின் தமிழ் தழுவிய கணினிசார் ஒலியனியலும், உரையிலிருந்து பேச்சாய்வும் (Ranganathan 2014, 2016) நல்ல நோக்கீடுகளாக அமையும்.

தமிழில் உரையிலிருந்து பேச்சு உருவாக்கத்திற்கான முன்னேற்பாடுகள் பல நடந்துள்ளன. குறிப்பாக சென்னையிலுள்ள IIT-யில் இது தொடர்பான ஆய்வுகள் மேற்கொள்ளப்பட்டுள்ளன. அந்நிறுவனத்தில் பணிபுரிந்த யக்ஙநாராயணா, ஹேமா மூர்த்தி, இராசேந்திரன் போன்றோர் உரையிலிருந்து பேச்சுத்தொடர்பான கருவிகளைச் செய்வதற்கு முயற்சிகள் மேற்கொண்டு வெற்றிபெற்றுள்ளனர். அண்ணாப்பல்கலைக்கழகத்தில் நடைபெற்ற RCTDIL-T என்ற ஆய்வுத்திட்டத்திலும் இதற்கான முயற்சிகள் செய்யப்பட்டுள்ளன. திருவனந்தபுரத்திலுள்ள உலகத் திராவிட மொழிகளின் பள்ளி (ISDL) மலையாளத்திற்கு உருவாக்கிய மலையாள உரையிலிருந்து பேச்சு ஒழுங்குமுறை தமிழ் உரைக்கும் நல்ல பேச்சு வெளியீட்டைத் தருகின்றது. எஸ் எஸ் என் கல்லூரியைச் சார்ந்த நாகராஜன் அவர்களும் உரையிலிருந்து பேசு உருவாக்கத்தில் தனது முத்திரையைப் பதித்துள்ளார்.

எதிரோலி என்பது தமிழிற்கான ஒரு உரையிலிருந்து பேச்சு இயந்திரம் ஆகும். உரையிலிருந்து பேச்சு கூறு தேவைப்படுகின்ற எந்தவொரு பயன்பாட்டிலும் செருக இயலுமாறு இயந்திரம் வடிவமைக்கப்பட்டுள்ளது. தமிழ் உரை உள்ளீடாக வழங்கப்படும்போது, அது முன் செயற்பாங்கு செய்யப்பட்டு, மொழியியல் விதிகளைப் பயன்படுத்தி ஒப்புருவுசார் மயக்கத்தை/தெளிவின்மைகளை அகற்றுவதற்காக ஒலியியல்சார் உருப்படுத்தம் செய்யப்படுகிறது. ஒரே எழுத்து அது நிகழும் நிலையைப் பொறுத்து வெவ்வேறு ஒலிகளைக் கொண்டிருக்கக்கூடும் என்பதால் ஒப்புருவுசார் மயக்கம்/தெளிவின்மை தமிழில் எழுகிறது. எடுத்துக்காட்டாக, தமிழில் உள்ள வல்லினம் எழுத்துக்கள் (க், ச், ட், த், ப், ற்,) ஒரு சொல்லுக்குள் நிகழும் நிலையின் அடிப்படையில் ஒன்றுக்கு மேற்பட்ட ஒலிகளைக் கொண்டுள்ளன. இந்த ஒலிபெயர்ப்பு உரையில் உள்ள ஒவ்வொரு வார்த்தையும் CVCV மாதிரியைப் பயன்படுத்தி அதனுடன் தொடர்புடைய ஒலியன்களாகப் பிரிக்கப்படுகிறது. ஒலியன்கள் அடையாளம் காணப்பட்டவுடன், ஒலியன்களுடன் தொடர்புடைய ஒலிக் கோப்புகள் ஒன்றிணைக்கப்பட்டு

பேச்சு வெளியீட்டைக் கொடுக்கும் வகையில் வரிசைப்படுத்தப்படுகின்றன. தற்போது, இந்த முறையைப் பயன்படுத்தி தமிழ் பேச்சை உருவாக்க 4000 அசைகள் பயன்படுத்தப்படுகின்றன.

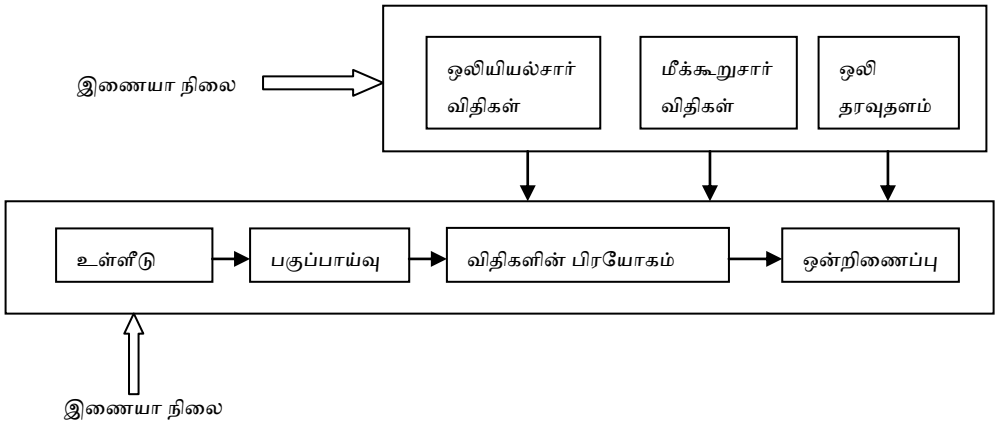
மேற்கண்ட உரையிலிருந்து பேச்சு மாற்ற ஒழுங்குமுறைகளில் ஒருசில விரிவாகத் கீழே தரப்பட்டுள்ளன.

3.5.8.2.1. பிலிப்ஸ் மற்றும் கூட்டாளிகளின் உரையிலிருந்து பேச்சு மாற்றம்

பிலிப்ஸ் மற்றும் பிறர் (Pilips et al 2005) 'உரையிலிருந்து பேச்சு மாற்றம்' என்ற கட்டுரையில் தாம் உருவாக்கியுள்ள உரையிலிருந்து பேச்சு மாற்ற ஒழுங்குமுறையைப் பற்றி விளக்குகின்றனர். பொதுவாக உரையிலிருந்து பேச்சு மாற்றம் மூன்று வழிகளில் செய்யப்படுகின்றது; அவை ஃபர்மண்ட்/ஒலிச்செறிவு அடிப்படை (formant based), அளவீடு அடிப்படை (parameter based), ஒன்றிணைப்பு அடிப்படை (concatenation based).

தற்போதைய உரையிலிருந்து பேச்சு மாற்ற ஒழுங்குமுறை இணைப்பாக்கம் அடிப்படையிலானதாகும். இணைப்பாக்கம் இரண்டு நிலைகளை உள்ளடக்கும்: இணையா நிலை (offline), இணை நிலை (online phase). இணையா நிலை அடிப்படை அலகு தெரிவு, மொழி விதிகளைக் கண்டறிதல் (ஒலியியல்சார் விதிகள் மற்றும் மீக்கூறுசார் விதிகள்) மற்றும் ஒலி தரவுத்தளத்தின் தெரிவு இவற்றை உட்படுத்தும். இணைநிலை உள்ளீட்டு உரையை அடிப்படை அலகுகளாகப் பிரித்தல் மற்றும் தமிழ் மொழி விதிகளைப் பிரயோகித்த பின்னர் அவற்றைப் பேச்சாக மாற்றுதல் இவற்றை உட்படுத்தும். இந்த ஒழுங்குமுறை ஒரு இடுகுறித்தன்மையான உரைக்கோப்பை ஏற்றுகொண்டு பொருளடக்கங்களை எழுத்துக்களாகப் பகுத்து உரை ஆய்வின் நிலைகளுக்கும் பகுபாய்வுக்கும் (அதாவது அடிப்படை அலகுகளின் கண்டறிதலுக்கு), மொழி விதிகளின் பிரயோகத்திற்கும் இறுதியாகப் பேச்சு வெளியீட்டை உருவாக்க ஒன்றிணைப்புக்கும் உருவாக்கத்திற்கும் கடத்துகின்றது.

படம்: உரையிலிருந்து பேச்சு எந்திரத்தின் தொகுதிப் படம்



மொழி விதிகள் (Language Rules)

தமிழ் எழுத்துக்கள்

தற்காலத் தமிழ் எழுத்து முறை 12 உயிரெழுத்துக்கள், 18 மெய் எழுத்துக்கள், 216 உயிர்மெய் எழுத்துக்கள் (அதாவது 18 × 12 = 216) மற்றும் "ஆய்தம்" எனப்படும் ஒரு எழுத்து ஆகியவற்றைக் கொண்டுள்ளது. ஆக மொத்தம் தமிழில் 247 மொத்தம் எழுத்துக்கள் உள்ளன.

உயிர்கள்

உயிர்கள் அவற்றின் உச்சரிப்புகளுடன் கீழ்வருமாறு தரப்பட்டுள்ளது.

உயிர்கள்	**உச்சரிப்புகள்**
அ	a
ஆ	aa/A
இ	i
ஈ	ii/I
உ	u
ஊ	uu/U
எ	e
ஏ	ee/E
ஐ	ai
ஒ	o
ஓ	ao/O

ஔ	au

மெய்கள் (Consonants)

மெய்கள் அவற்றின் உச்சரிப்புகளுடன் கீழ்வருமாறு தரப்பட்டுள்ளது.

மெய்கள்	**உச்சரிப்புகள்**
க்	k
ங்	ng
ச்	ch
ஞ்	nj
ட்	t
ண்	N
த்	th
ந்	nth
ப்	p
ம்	m
ய்	y
ர்	r
ல்	l
வ்	v
ழ்	zh
ள்	L
ற்	tR
ன்	n

இந்த மெய்கள் மூன்றாகப் பகுக்கப்பட்டுள்ளன.

வல்லினம்: க், ச், ட், த், ப், ற்

மெல்லினம்: ங், ஞ், ண், ந், ம், ன்

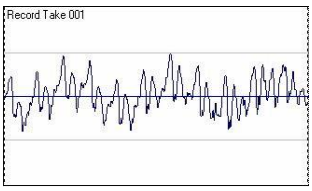
இடையினம்: ய், ர், ல், வ், ழ், ள்

ஆயுதம்: ஆயுத எழுத்து ஃ என்று எழுதப்படும்; h என்று உச்சரிக்கப்படும். இது தமிழில் குறைவான சொற்களிலேயே பயன்படுத்தப்படுகின்றது.

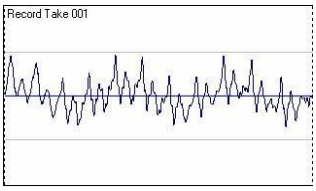
உயிர்மெய் எழுத்துக்கள்: 18 மெய் எழுத்துக்கள் 12 உயிர் எழுத்துக்களுடன் சேர்ந்து வருகையில் 18x12 = 216 உயிர்மெய் எழுத்துக்கள் விளைகின்றன.

தமிழ் ஒலியியல் சார் விதிகள் (Tamil phonetic rules)

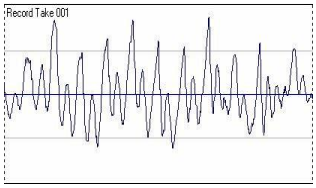
தமிழ் எழுத்துகள் தனித்தன்மையான ஒலிகளுள்ள வடிவங்களைக் கொண்டுள்ளன. சில எழுத்துக்கள் அவை வரும் இடம் அடிப்படையில் வேறுபட்ட ஒலிகளைக் கொண்ட ஒரே வடிவைக் கொண்டிருக்கின்றன. எனவே இந்த இடம் அடிப்படையிலான ஒலிப்பு வேறுபாடு விளக்கப்படவேண்டும். எடுத்துக்காட்டாக *க* என்ற எழுத்து அவை வரும் இடம் அடிப்படையில் [ka], [ha], [ga] என்ற ஒலிப்பு வேறுபாடுகளைக் கொண்டுள்ளது.



க (ka)



க (ha)



க (ga)

தமிழ் வல்லின எழுத்துகள் க், ச், ட், த், ப், ற் என்பன அவை வரும் இடம் அடிப்படையில் வேறுபட்டு ஒலிக்கும்.

இயல்பு ஒலி (Normal sound): வல்லின மெய்களின் முன் அவைகளே வரும்போதும் மற்றும் சொல்லின் தொடக்கத்திலும் மாற்றமின்றி இயல்பு ஒலிகளாக (குரலிலா ஒலிகளாக/voiceless sounds) வரும்.

தமிழ் எழுத்து	உச்சரிப்பு
க	ka
ச	cha
ட	ta
த	tha
ப	pa
ற	tRa

கடின ஒலிகள் (Hard Sounds): வல்லின மெய்கள் முன் மெல்லினங்கள் ங், ஞ், ண், ந், ம், ன் வருகையில் அவை கடின ஒலிகளாக (குரல் ஒலிகளாக/voiced sounds) வரும்.

தமிழ் எழுத்து	உச்சரிப்பு	முன்னர்வரும் ஒலி
க	ga	ங்
ச	cha	ஞ்
ட	da	ண்
த	dha	ந்
ப	bha	ம், ன், ண்
ற	dRa	ன்

மெல் ஒலி (Soft Sound): மெல் ஒலி சொல்லின் மத்தியில் அவற்றின் மெய்கள் முன்னர் வராத நிலையில் வரும்.

தமிழ் எழுத்து	உச்சரிப்பு
க	ha
ச	sa
ட	da
த	dha
ப	bha
ற	Ra

தமிழ் மீக்கூறு விதிகள் (Tamil Prosodic rules)

மீக்கூறு என்பது வெளிப்பாட்டிற்கு பயன்படுத்தப்படுகிற ஒரு பௌதிக ஒலியனியல்சார் விளைவு ஆகும். இந்த இயற்கையான ஒலி உரையை உருவாக்கப் பயன்படுகிறது. மீக்கூறு முறையான புலணர்வைக் கொண்டுள்ளது; மற்றும் இடைநிறுத்தங்கள், இசைமை, காலம் மற்றும் சத்தத்தின் அடிப்படையில் பேச்சாளரின் நோக்கங்களை மீட்டெடுப்பதையும் கொண்டுள்ளது. பேச்சு சமிக்ஞையின் மிக வெளிப்படையான பகுதி இசைமை ஆகும். மக்கள் தங்கள் இசைமை மாறுபாடுகள் மூலம் தங்கள் உணர்ச்சிகளை வெளிப்படுத்த முயற்சிக்கிறார்கள். எனவே இயற்கையான ஒலிப் பேச்சைப் பெற இசைமை மற்றும் கால அளவு பற்றிய விரிவான பகுப்பாய்வு செய்யப்பட வேண்டும். இது பல மறு செய்கைகளுடன் கூடிய மிகவும் கடினமான பணி என்பதால், இந்தச் செயல்பாட்டின் ஒவ்வொரு அலகின் காலமும் மட்டுமே கருதப்பட்டுள்ளது. தேவைப்படும்போது சரியான அளவு மௌனத்தை (silence) செருகுவதும் ஒரு முக்கியமான பணியாகும். நிறுத்தற்குறி/முழுப்புள்ளி (full-stop), கால்புள்ளி (coma), அரைப்புள்ளி

(semicolon), முக்காற்புள்ளி (colon) போன்ற இடங்களில் அறிமுகப்படுத்தப்பட வேண்டிய மாறுபாடுகள் கணக்கில் எடுத்துக்கொள்ளப்பட்டது.

நடைமுறைப் படுத்துதல்/செயல்படுத்துதல் (Implementation)

இணையா (ஆஃப்லைன்) கட்டம், இணை (ஆன்லைன்) கட்டம் மற்றும் பயனர் இடைமுகத்தின் உருவாக்கம் இவற்றை உள்ளடக்கிய உரையிலிருந்து பேச்சு இயந்திரத்தை செயல்படுத்துவது இந்த பிரிவில் விவரிக்கப்படுகிறது.

இணையாக் (ஆஃப்லைன்) கட்டம்

கணினியின் இணையா (ஆஃப்லைன்) செயல்பாட்டில் அடிப்படை அலகுகளைத் தேர்ந்தெடுப்பது மற்றும் ஒலி தரவுத்தளம் உருவாக்குவது ஆகியவை அடங்கும்

அடிப்படை அலகின் தெரிவு (Selection of basic unit)

அடிப்படை அலகுகளின் தொகுப்பில் தமிழில் உள்ள அனைத்து ஒலிகளும் / சொற்களும் இருக்க வேண்டும். பேச்சின் இந்த அலகு அசைகள், ஒலியன்கள், ஈரொலிகள், சொற்கள், சொற்றொடர்கள் அல்லது வாக்கியங்களாக இருக்கலாம். இவற்றில் மிக நீளமான அலகு வாக்கியம் ஆகும்; மற்றும் குறுகிய அலகு ஒலியன்/ஈரொலி ஆகும்.

நாம் வாக்கியத்தை (sentence) அடிப்படை அலகு எனத் தேர்வுசெய்தால், ஒன்றிணைப்புகளின் எண்ணிக்கை குறைவாக இருக்கும். எனவே தரம் அதிகமாக இருக்கும். ஆனால் அலகுகளின் எண்ணிக்கை எல்லையற்றதாக இருக்கும், அது இருக்கும் எந்தவொரு ஸ்கிரிப்டுடனும் வேலை செய்ய அத்தகைய அமைப்பை வடிவமைக்க இயலாது.

Diphone

நாம் ஒலியனை (phoneme) அடிப்படை அலகு எனத் தெரிவுசெய்தால், ஒன்றிணைப்புகளின் எண்ணிக்கை அதிகமாக இருக்கும். எனவே தரம் குறைவாக இருக்கும். இருப்பினும், தரவுத்தளத்தின் அளவும் ஒரு முக்கியமான கருத்தில் கொள்ள வேண்டிய காரணி ஆகும்.

அசை (syllable) அடிப்படை அலகு எனத் தேர்ந்தெடுக்கப்பட்டால், அசை வெவ்வேறு இடங்களில் வரும்போது அவை ஒவ்வொன்றின் ஒலி மாறுபாடுகளையும் செயல்படுத்த வேண்டும்; அது சாத்தியமில்லை; ஏனென்றால் ஒரு அசை தமிழ் எழுத்துக்களை மட்டுமே கொண்டுள்ளது.

ஈரொலிகள் (டிஃபோன்/diphone) என்பது இரண்டு அருகிலுள்ள இரண்டு அசைகளுக்கு நடுவில் தோன்றும் ஒலி ஆகும். என்றால் இது அடிப்படை அலகு எனத் தேர்ந்தெடுக்கப்பட்டால் ஒன்றிணைப்பு உருக்குலைவு குறைவாக இருக்கும். ஆனால் தரவுத்தளத்தில் சேமிக்க வேண்டிய

அலகுகள் எண்ணிக்கை அதிகம் ஆகும்; ஏனென்றால் ஒவ்வொரு அருகிலுள்ள அசைகளின் ஒலிகளை சேமிக்க வேண்டியது அவசியம் ஆகும்.

இந்த உண்மைகள் அனைத்தையும் பரிசீலித்தபின், ஒலியன் இதற்கான அடிப்படை அலகு எனத் தேர்ந்தெடுக்கப்பட்டுள்ளது. இது ஒரு மொழியின் அடிப்படை ஒலி அலகு ஆகும். ஆங்கிலத்தைப் பொறுத்தவரை, 26 எழுத்துக்கள் உள்ளன; ஆனால் ஆங்கிலத்தில் நிறைய மயக்கங்கள்/தெளிவற்ற தன்மைகள் உள்ளன. தமிழில் மிகவும் குறைவான மயக்கமே/தெளிவின்மை உள்ளது அவற்றில் மிகவும் தனித்துவமான உச்சரிப்புகளுடன் 247 உள்ளன உள்ளன என்பதே உண்மை. தி பல்வேறு எழுத்துகள் தமிழில் சுமார் 343 ஒலியன்களின் பெரிய பட்டியலை உருவாக்குகின்றன, இது தமிழ் பேச்சுக்கு அடிப்படையாக அமைகிறது.

ஒலி தரவுத்தளத்தை உருவாக்குதல் (Creation of sound database)

ஒலியன்கள் அடிப்படை அலகு எனத் தேர்ந்தெடுக்கப்பட்டதால், எல்லா ஒலியன்களும் பதிவு செய்யப்பட்டு தரவுத்தளத்தில் சேமிக்கப்படுகிறது. ஒலி தரவுத்தளத்தின் தரம் ஒரு முக்கியமான காரணியாகும் உரையிலிருந்து பேச்சைப் பொருத்தவரையில் வெளியீட்டின் தரம் முக்கியமானதாகும்.

பேச்சைப் பதிவுசெய்யும்போது, குறைந்தது 8 கிலோஹெர்ட்ஸ் பதமாதிரி (sampling) வீதமும் 8-பிட்டும் பதமாதிரியும் இருக்க வேண்டும். இது மோசமான தரமான பேச்சை உருவாக்குகிறது, ஆனால் அதைப் புரிந்து கொள்ள முடியும். தரத்தை மாதிரியில் உள்ள பிட்களின் எண்ணிக்கையை 12 பிட்கள் அல்லது 16 பிட்களாக அதிகரிப்பதன் மூலம் மேம்படுத்தலாம். டிஜிட்டலாக்க நிலைகளை அதிகரிப்பது ஒலி அடையாளக்குறியை சத்த விகிதத்திற்கு அதிகரிக்கும். பதமாதிரி விகிததை 8 kHz க்கு மேல், அதாவது 10 kHz, 16 kHz அல்லது 20 kHz வரை அதிகரிப்பது அதிர்வெண் உள்ளடக்கத்தை மேம்படுத்துகிறது. அதிக பதமாதிரி அதிர்வெண் அதிக தரத்தைத் தரும். உயர்தரப் பேச்சு பதிவுக்குக்கும் திரும்பி கேட்பதற்கும் 16 கிலோஹெர்ட்ஸ் பதமாதிரி வீதம் நியாமானதாகும்.

ஒலி தரவுத்தளத்தை செயல்படுத்துவது 4 கட்டங்களில் மேற்கொள்ளப்பட்டது.

1. ஒலிகளைப் பதிவுசெய்தல் (Recording sounds): ஒலிகளைப் பதிவு செய்வது ஒரு இயல்பு தமிழ்மொழி பேசுபவரைப் பயன்படுத்தி மேற்கொள்ளப்பட்டது. 30-16000 ஹெர்ட்ஸ் அதிர்வெண் பதில் கொண்ட மைக்ரோஃபோன் மூலம் சத்தம் இல்லாத அறையில் இது செய்யப்பட்டது. 16 பிட்கள் பதமாதிரி அளவுடனும் 8 kHz பதமாதிரி அதிர்வெண்ணுடனும் பதிவு செய்யப்பட்டது.

பதமாதிரி அதிர்வெண் சிறியது என்றாலும் புரிந்துகொள்ளக்கூடிய குரல் வெளியீட்டை உருவாக்க இது போதுமானதாக இருந்தது.

2. ஒலியன்களைப் பிரித்தெடுத்தல் (Extracting phonemes): பதிவுசெய்யப்பட்ட ஒலி கோப்பு அனைத்து அலகுகளும் (உயிரெழுத்துகள், மெய் மற்றும் மெய்-உயிர் பிரிவுகள்) தொடர்ச்சியான கோப்பாகும். ஒவ்வொரு ஒலியனுக்கும் பொருந்தும் பகுதி இந்த கோப்பிலிருந்து பிரித்தெடுக்கப்பட்டு ஒரு தனி *.wav கோப்பாக சேமிக்கப்பட்டது. இது ஒலி எடிட்டிங் மென்பொருளான “Sound Forge”ஐப் பயன்படுத்தி செய்யப்பட்டது.

3. இசைமை குறித்தல் (Pitch marking): இதன் பொருள் ஒவ்வொரு ஒலியனின் உண்மையான தொடக்க மற்றும் இறுதி புள்ளியைக் கண்டறிதல் ஆகும். இசைமை குறித்தல்களில் அலைவடிவங்கள் ஒன்றிணைக்கப்படுவதால் இது அவசியம் ஆகும். எனவே ஒவ்வொரு ஒலியனின் தொடக்கமும் முடிவும் கவனமாகக் கேட்பதன் மூலமும் கண்டறியப்பட்டது மற்றும் அந்த இரண்டிற்கும் இடையே உள்ள பகுதி தேர்ந்தெடுக்கப்பட்டு ஒரு தனி *.wav கோப்பாக சேமிக்கப்பட்டது. அதன்பிறகு தி ஒவ்வொரு ஒலியனின் உண்மையான கால அளவும் “சோதனை மற்றும் பிழை” (“trial and error”) மூலம் தீர்மானிக்கப்பட்டது.

4. ஒலி கோப்புகளைத் திருத்துதல் (Editing sound files): இசைமை குறிக்கப்பட்ட ஒலி கோப்புகள் பெரும்பாலான சந்தர்ப்பங்களில் ஏற்றுக்கொள்ளக்கூடிய தரத்தில் இருபதில்லை. எனவே இணைத்தால் அவை ஒரு நல்ல தரமான வெளியீட்டை உருவாக்கும்படி அவற்றின் தரத்தை மேம்படுத்த திருத்த வேண்டும். இந்த கட்டத்தின் போது சத்தம் குறைப்பு வசதிகள் உள்ள “சவுண்ட் ஃபோர்ஜ்” என்பதும் பயன்படுத்தப்பட்டது.

இணைப்புக் (ஆன்லைன்) கட்டம்

இணைப்பு/ஆன்லைன் கட்டத்தில் நிரல் இருக்கும் போது பின்னணியில் மேற்கொள்ளப்படும் பணிகள் அடங்கும் இயங்குகிறது. இது மூன்று முக்கியமான கட்டங்களை உள்ளடக்கியது, அதாவது உரை பகுப்பாய்வு மற்றும் பாகுபடுத்தல் (text analysis & parsing), மொழி விதிகளின் பிரயோகம் மற்றும் உரையிலிருந்து பேச்சு மாற்றம் உருவாக்க மொழி அலகுகளின் ஒருங்கிணைப்பு என்பன.

உரை பகுப்பாய்வு மற்றும் பாகுப்பாய்வு: இந்த கட்டம் உள்ளீட்டு உரையை பகுப்பாய்வு செய்து ஒற்றை கிளிஃபைக் எழுத்துக்கள், இரட்டை கிளிஃப் எழுத்துக்கள் மற்றும் மூன்று கிளிஃப் எழுத்துக்கள் காண்டுபிடிக்கின்றது. பின்னர் அது உள்ளீட்டு உரையைப் பேச்சின் அடிப்படை அலகுகளின் வரிசையாகப் பிரிக்கின்றது.

மொழி விதியின் பிரயோகம்/பயன்பாடு: உள்ளீட்டு உரையை அதன் அடிப்படை அலகுகளாகப் பிரித்த பிறகு, தரவுத்தளத்திலிருந்து சரியான ஒலியனைப் பிரித்தெடுக்க மொழி விதிகள் பயன்படுத்தப்படுகின்றன. இந்த கட்டத்தில் இயற்கையான ஒலிப்புப் பேச்சை ஒருங்கிணைப்பதற்காக/உருவாக்குவதற்காக மொழி திட்டவட்டமான ஒலிப்பு விதிகள் இணைத்துக்கொள்ளப்பட்டன.

ஒன்றிணைப்பு: இறுதி கட்டத்தின் போது, சரியான வரிசையில் அடிப்படை அலகுகளின் அலைவடிவங்களின் ஒன்றிணைப்பு தேவையான ஒலி அலைவடிவத்தை உருவாக்க மேற்கொள்ளப்படுகின்றது. அடிப்படை அலகுகளின் அலை வடிவங்கள் தரவுத்தளத்திலிருந்து பிரித்தெடுக்கப்பட்டு ஒன்றிணைக்கப்படுகின்றன.

பயனர் இடைமுகம் (User Interface)

விஷுவல் பேசிக் 6.0-ஐப் பயன்படுத்தி ஒரு எளிய வரைகலை பயனர் இடைமுகம் graphical user interface (GUI) உரையிலிருந்து பேசு இயந்திரம் உருவாக்கப்பட்டது. இடைமுகம் தமிழில் உரையை உள்ளிட உரைப் பெட்டியையும் பேச்சு ஒலியை இயக்க பொத்தானையும் கொண்டுள்ளது. உள்ளீட்டு உரை தமிழ் வார்த்தையாகவோ ஒரு எண்ணாகவோ வாக்கியமாகவோ இருக்கலாம். உரையிலிருந்து பேச்சு மாற்றத்தை உருவாக்க உள்ளிடப்பட்ட உரையை சேமித்துவைக்க அல்லது முன் சேமித்துவைத்த உரையைத் திறக்க ஒரு விருப்பத்தை இடைமுகம் பயனருக்கு வழங்குகிறது. “TSC_Avarangal” என்ற தமிழ் எழுத்துரு பயனர் இடைமுகத்தின் உருவாக்கத்தில் பயன்படுத்தப்பட்டது. “eKalappai 1.0” என்ற மென்பொருள் தமிழ் உரையை உள்ளீடு செய்ய பயன்படுத்தப்பட்டது.

இதைப்பற்றி பின்வருமாறு முடிவுரை தரப்பட்டுள்ளது:

எந்தவொரு தன்னிச்சையான தமிழ் உரையையும் பேச்சுக் கூற்றாக மாற்றக்கூடிய ஒரு உரையிலிருந்து பேசு மாற்று இயந்திரத்தை தமிழ் மொழிக்கு வெற்றிகரமாக உருவாக இயலும் என்பதை இந்தச் செயல்பாடு காட்டுகிறது. தமிழ் மொழியின் அடிப்படை அலகுகளாகத் தேர்ந்தெடுக்கப்பட்ட 343 ஒலியன்களால் உருவாக்கப்பட்ட இயந்திரத்தின் பின்தளம் ஒழுங்கமைப்பின் இதயமாக செயல்படுகிறது. “Sound Forge” என்ற ஒலி எடிட்டிங் மென்பொருள் ஒலியன்களை சரியான இசைமையுடன் தயாரிக்கும் போது பயன்படுத்தப்பட்டது. ஒலியன்களைப் பதிவுசெய்ய 8 kHzஐ விட சிறந்த பதமாதிரி வேகத்தைப் பயன்படுத்துவதன் மூலம் ஒலி வெளியீட்டின் தரம் மேலும் அதிகரிக்க முடியும். உருவாக்கப்பட்ட உரையிலிருந்து-பேச்சு இயந்திரம் கொடுக்கப்பட்ட உரையை அடிப்படை அலகுகளாகப் பிரிப்பது, சரியான

ஒலியன்களைப் பிரித்தெடுப்பது, தரவுத்தளத்துடன் பொருந்துவது மற்றும் தேவையான ஒலி அலைவடிவத்தை உருவாக்குவதற்காகச் சரியான வரிசையில் அடிப்படை அலகு அலைவடிவங்களை ஒன்றிணைப்பது என்பனவற்றை வெற்றிகரமாகச் செய்தது. என்ஜினின் முன் இறுதியில் விஷுவல் பேசிக் மூலம் உருவாக்கப்பட்ட ஒரு வரைகலை பயனர்கள் இடைமுகம் ஆகும். இது பயனர்கள் தமிழ் உரையை எழுத்துக்களில் அல்லது எண்களில் உள்ளிட அனுமதிக்கிறது. இந்த இடைமுகம் பயனர்களை உரையை பின்னர் சேமித்து மீண்டும் இயக்க அனுமதிக்கிறது. பயனர் இடைமுகத்தை உருவாக்க எழுத்துருவாக “TSC_Avarangal” தமிழ் எழுத்துரு தேர்ந்தெடுக்கப்பட்டது.

இந்த ஆய்வில் சரியான ஒன்றிணைப்பிற்காக அடிப்படை ஒலி அலகின் ஒலிப்பு காலத்தைச் சரிசெய்ய முக்கியத்துவம் கொடுக்கப்பட்டது. இருப்பினும், வெளியீட்டை இயற்கையான பேச்சுக்கு நெருக்கமாக மாற்றுவதற்கான இசைமையைச் சரிசெய்தல் மூலம் மேலும் மேம்பாடுகள் மேற்கொள்ளப்பட வேண்டும்.

3.5.8.2.2. உதயகுமார் மற்றும் கூட்டாளிகளின் தமிழிற்கான எழுத்திலிருந்து ஒலியனாக மாற்றும் தீர்மானக் கிளையமைப்புக் கற்றல் (Decision Tree Learning for Automatic Grapheme to Phoneme Conversion for Tamil)

’தமிழிற்கான எழுத்திலிருந்து ஒலியனாக மாற்றும் தீர்மானக் கிளையமைப்புக் கற்றல்’ உதயகுமாரும் கூட்டாளிகளும் (Udhyakumar et la 2004) தீர்மானக் கிளையமைப்பு கற்றலைப் பயன்படுத்தி தமிழ் எழுத்தை ஒலியனாக ஒரு புதிய அணுகுமுறையை விளக்குகின்றனர். முன்மொழியப்பட்ட இந்த அணுகுமுறை, விதி அடிப்படையிலான அண்குமுறை போலன்றி பரந்த அளவிலான சூழல் விதிகளை உருவாக்க முடியும்; இதனால் மாற்றத்திற்குச் சிறந்த துல்லியத்தைத் தரும்.

எழுத்திலிருந்து ஒலியன் மாற்றம் (Grapheme to phoneme conversion) உரையிலிருந்து பேச்சு மாற்ற அமைப்புக்கும் பேச்சு அறிதல் அமைப்புக்கும் முக்கியமானது ஆகும். மொழியியலாளர்களால் தயாரிக்கப்பட்ட அகராதிகள் மிகவும் நம்பகமான ஆதாரமாக இருந்தாலும், அவை எந்த மொழியிலும் உள்ள சொற்களின் முழுமையான பட்டியலை உட்படுத்த முடியாததால் அவை ஒருபோதும் முழுமையானவை அல்ல. ஒரு விரிவான அகராதி நிறைய சேமிப்பு இடத்தையும் பயன்படுத்துகிறது. இது பேச்சு தொழில்நுட்பங்களை மாறும் சொல்லகராதி சூழல்களுக்கு பயன்படுத்துவதை கட்டுப்படுத்துகிறது. எனவே எழுத்து வடிவத்திலிருந்து உச்சரிப்பைக் கணிக்கும் மாற்று நடைமுறை அவசியம் ஆகும்.

தானியக்கமாக எழுத்தை ஒலியனாக மாற்றுவதற்கான பல அணுகுமுறைகள் பயன்படுத்தப்பட்டுள்ளன. ஒலிப்பு மொழிகளால் ஈர்க்கப்பட்டு, எழுத்தை ஒலியனாக மாற்றுவது பொதுவாக ஒரு குழும விதிகளால் மேற்கொள்ளப்படுகிறது. அவைகளை உருவாக்க நிறைய நிபுணத்துவம் தேவைப்படுகிறது மற்றும் சில நேரங்களில் விதி குறுக்கீடு மற்றும் மிகையைச் சரிபார்க்க கடினமாக உள்ளது. ஆங்கிலம், பிரஞ்சு மற்றும் ஜெர்மன் போன்ற பல மொழிகளில் எழுத்தை ஒலியனாக மாற்றும் விதிகளுக்கு புள்ளிவிவர மாதிரியமைத்தல் (Statistical modeling) வெற்றிகரமாக பயன்படுத்தப்பட்டுள்ளது. எழுத்தை ஒலியனாக மாற்றுவதற்கு நரம்பியல் வலைப்பின்னலின் (neural networks) பயன்பாடு ஆராயப்படுகிறது. இந்த அணுகுமுறைகள் விதி எழுதும் நேரத்தைத் தவிர்க்கும்; எனவே பயிற்சிச் செயல்முறையையை முழுவதும் தானியக்கமாகச் செய்ய இயலும். எழுத்திலிருந்து ஒலியன் பொருத்தத்திற்கான ஒரு தீர்மானக் கிளை சார்ந்த தானியங்கி இக்கட்டுரையில் ஆய்வு செய்யப்படுகிறது. இதன் முக்கிய நோக்கம், தமிழ் உரையிலிருந்து பேச்சு மாற்றம் மற்றும் பேச்சு அறிதல் (பேசிலிருந்து உரைக்கு மாற்றம்) ஆகிய பயன்பாடுகளுக்காக தீர்மானக் கிளை அணுகுமுறையைப் பயன்படுத்தி ஒரு தானியங்கி ஒழுங்குமுறைய உருவாக்குவது ஆகும்.

பிற இந்திய மொழிகளைப் போலவே தமிழும் எழுத்துக்கும் ஒலியனுக்கும் இடையில் கிட்டத்தட்ட ஒன்றுக்கு பொருத்தத்தைக் காட்டுகின்றது. இருப்பினும் நேரடியான எழுத்திலிருந்து ஒலியன் பொருத்தம் சாத்தியம் இல்லை; ஏனென்றால் தமிழ் எழுதுக்கள் சில சூழல்களில் வேறுபட்ட ஒலியை உருவாக்கும். எடுத்துக்காட்டாக க் என்பது கடல் என்பதில் k எனவும் தங்கம் என்பதில் g எனவும் பகல் என்பதில் h எனவும் உச்சரிக்கப்படும். தமிழில் உருபனியல் மற்றும் ஒலியியல் ஆய்வுகள் இந்த வேறுபாடுகள் சில சிக்கலான மொழிச் சீர்மைகளைப் (complex language regularities) பின்பற்றி அமைகின்றன என நிரூபிக்கின்றன. இந்த சீரான தன்மை தீர்மான கிளையமைப்புகள் போன்ற (decision trees) தரவு-இயக்க நுட்பங்களைப் (data-driven technique) பயன்படுத்தத் தூண்டுகிறது. கிளையமைபுகள் மனித முயற்சியால் ஒலிபெயர்ப்பு செய்யப்பட்ட மிகப் பெரிய அகராதியால் பயிற்சி செய்யப்படவேண்டும். தமிழுக்கு அத்தகைய உச்சரிப்பு அகராதி இல்லை. எனவே உச்சரிப்பு அகராதியை உருவாக்க விதி அடிப்படையிலான ஒழுங்குமுறை இங்கு பயன்படுத்தப் படுகின்றது; இது பின்னர் கையால் திருத்தப்படும்.

கட்டுரை பின்வருமாறு அமையும். பிரிவு 1 அறிமுகம் ஆகும்; பிரிவு 2 தமிழ் எழுத்துமுறை ஒலியன் தொகுப்பு பற்றி விவரிக்கிறது. விதி அடிப்படையிலான மற்றும் தீர்மான கிளையமைப்பு அணுகுமுறைகள் பிரிவு 3 மற்றும் 4இல் முறையே விளக்கப்பட்டுள்ளது. பிரிவு 5 தீர்மானக்

கிளையமப்பை பயன்படுத்தி தானியமாக எழுத்திலிருந்து ஒலியன் மாற்றுவதன் முடிவுகளை வழங்குகிறது. இறுதி பிரிவு 6 முடிவுரையாகும்.

தமிழ் எழுத்துமுறையும் ஒலித் தொகுதியும்

தமிழ் எழுத்து முறை 12 உயிரெழுத்துக்களைக்களையும் 18 மெய் எழுத்துக்களையும் கொண்டுள்ளது. மெய் எழுத்துக்கள் தனித்தனியாக உயிரெழுத்துக்களுடன் இணைந்து உயிர்மெய் என்ற 216 கூடுதல் எழுத்துக்களின் புதிய தொகுப்பை உருவாக்கும். ஃ என்ற எழுத்துக்கு ஆயுதம் என்ற தனித்துவமான பெயர் உண்டு. தமிழ் ஒலிகள் 43 ஒலியன்களாக வகைப்படுத்தப்பட்டுள்ளன. அட்டவணை 1 தமிழ் ஒலியன் குழுமத்தின் விவரங்களை அளிக்கிறது.

எழுத்து	ஒலி	எடுத்துக்காட்டு
அ	ax	அம்மா
ஆ	aa	ஆடு
இ	ih	இலை
ஈ	iy	ஈகை
உ	uh	உணவு
	U	
ஊ	uw	தப்பு
எ	ae	எட்டு
ஏ	ey	ஏணி
ஒ	ao	ஒன்று
ஓ	ow	ஓம்
ஔ	aw	ஔ
ஷ்	sh	விஷம்
ஜ்	jh	ஜாதி
க்	k	கல்வி
	g	பங்கு
	hh	பகல்
ங்	ng	தங்கம்
ச்	ch	காட்சி

	s	சங்கு
ஞ்	ny	ஞானம்
ட்	t	அட்டை
	d	வண்டி
ண்	nn	கண்
	tx	தட்டு
	dh	பந்தம்
	T	மதி
ந்	nd	பந்து
	n	நான்
ப்	p	பாடம்
	b	அம்பு
	P	லாபம்
ம்	m	கரும்பு
ய்	y	யமுனை
ர்	r	மரம்
ல்	l	மாலை
வ்	v	வயது
ழ்	lzh	பழம்
ள்	ll	உள்ளம்
ற்	tr	ஒற்றன்
	dr	கன்று
	rr	அறம்
ன்	n	அன்னை

3.5.8.2.2.3. விதி அடிப்படையிலான எழுத்திலிருந்து ஒலியன் மாற்றம் (Rule-based Grapheme to phoneme conversion)

அட்டவணை 1இல் உள்ள எடுத்துக்காட்டுகளிலிருந்து தமிழ் எழுத்துக்களுக்கும் ஒலியன்களுக்கும் இடையே ஒன்றுக்கு ஒன்று ஒரு சிறிய வரைபடம் இல்லை அது தெளிவாகிறது..

இருப்பினும், பெரும்பாலான நேர்வுகளில் இந்த மாறுபாடுகள் சூழலின் அடிப்படையில் குறியத்திறவு செய்யப்படலாம். எனவே வலுவுள்ள சூழல் சார் விதிகள் உச்சரிப்பைத் துல்லியமாக கணிக்க வடிவமைக்கப்பட்டுள்ளது. விதிகள் பார்வை, சூழல் மற்றும் இலக்கு ஒலியன் குறித்த குறிப்பிட்ட மரபுகளைக் கொண்டுள்ளன. எ.கா. விதி மூச்சொலி k (*) ⇒ g என்பது k முன்னர் மூச்சொலியும் (எ.கா. ng (ங்)) வேறு ஏது ஒலியாலும் தொடர்ப்படுகையில் பங்கு என்பதில் உள்ளது போல் g ஆக மாறும். எந்த வல்லுநர் அமைப்போலவும் எல்லாத் தொடர்புடைய நேர்வுகளை எதிர்பார்ப்பது கடினமாகும். எடுத்துக்காட்டாக கிராமம் என்பதில் உள்ள k என்பது மூக்கொலி முன் வராமலேயே g என உச்சரிக்கப்படும். இதுபோன்று பந்தம் என்பதிலுள்ள p என்பது மூக்கொலி முன் வராமலேயே b என உச்சரிக்கப்படும். இந்த விதிவிலக்குகள் தனியாக விதிவிலக்கு அகராதியில் சேகரிக்கப்படவேண்டும். பொதுவாக விதிவிலக்கு அகராதிக்கும் விதிகளின் கலவைத்தன்மைக்கும் இடையில் ஈடுகட்டல் இருக்கும். இது எழுத்திலுருந்து ஒலியன் பொருத்தம் மிகச் சிக்கலானது என்பதை நிரூபிக்கும்; இது தீர்மான கிளையமைப்பு போன்ற எந்திரம் கற்றல் வழிமுறைகளுக்கு (machine learning algorithms) இந்த புலத்தைக் கவர்சிகரமாய் ஆக்குகின்றது.

3.5.8.2.3. தீர்மானக் கிளையமைப்பு அடிப்படையிலான எழுத்திலிருந்து ஒலியன் மாற்றம் (Decision tree based Grapheme to phoneme conversion)

அண்டை ஒலிகளின் இயல்பு அடிப்படையில் ஒரு ஒலியனின் ஒலியியல் மாறுபாடுகளைக் கண்டுபிடிக்க தீர்மான கிளையமைப்புகள் பயன்படுத்தப்படலாம். இதை எழுத்து வடிவில் அடுத்துவருகிற எழுதுகளுக்கு ஏற்ப விளக்கலாம்; இதையொட்டி இந்த ஒழுங்குமுறை எழுத்துகளை ஒலிபெயர்ப்பை உருவாக்க இந்த ஒலியியல் வேறுபாடுகளை பயன்படுத்த இயலும்.

தீர்மானக் கிளையமைப்புகள் விதிகளையும் முனைகளையும் கொண்ட கணுக்களைக் கொண்டிருக்கும்; அவை இலக்கு வகுப்புகளுடன் புலக்குறிப்பு செய்யப்பட்டிருக்கும். ஒவ்வொரு கணுக்களிலும் ஒரு தீர்மான அடிப்படை பொருளை இடது அல்லது வலது துணைக் கிளையமைப்புக்கு ஒதுக்கும். பொருள் ஒரு இலையை அடையும் போது, இந்த இலையின் வகுப்பு புலக்குறிப்பு குறிப்பிட்ட எழுதுப்பெயர்ப்புக்குப் பதிலாகப் பயன்படுத்தப்படுகிறது.

வரிசைப்படுத்தும் செயல்முறை

பயிற்சிக்கு முன், ஒவ்வொரு எழுத்தும் அதனுடன் தொடர்புடைய ஒலியுடன் வரிசைப்படுத்தப்படவேண்டும். தமிழில் எழுத்துக்களுக்கும் ஒலிகளுக்கும் இடையே உள்ள வலுவான உறவு காரணமாக, சிக்கலான வரிசைப்படுத்தும் நடைமுறைகள் தேவையில்லை.

தமிழில் எழுத்துக் கோர்வைகள் ஒலி வரிசைகளுக்குச் சமமாகவோ அதைவிட நீளமாகவோ இருக்கும். ஒலி வரிசைகளுடன் ஒன்றுக்கொன்றான வரிசைப் பொருத்தத்தைக் கொண்டிருக்க வேண்டி) பொருத்தமான இடங்களில் ஒலியனியல்சார் எப்சிலோன்கள் (பூஜ்யங்கள் செருகப்படுகின்றன. எடுத்துக்காட்டாக *சென்னை* என்பதிலுள்ள எழுத்து ச் /ch/ எனவோ /s/ எனவோ உச்சரிக்கப்படலாம். நாம் ஒலிக் குழுமத்தில் /ch*s/ என்ற புதிய நுழைவைச் சேர்ப்பதன் மூலம் இந்த நேர்வைச் சமாளிக்கலாம். தொடர்புடைய ஒலியுடன் பொருத்த அனுமதிக்கப்பட்ட ஒலிகளின் தொகுப்பு ஒவ்வொரு எழுத்துக்கும் குறிப்பிடப்பட்டுள்ளது.

தீர்மானக் கிளையமைப்புகளின் பயிற்சி (Training of decision trees)

அடிப்படை தீர்மானக் கிளையமைப்பு ஆம்-இல்லை என்ற கேள்விகளின் தொகுப்பையும் மற்றும் வேரிலிருந்து கிளையமைப்பை வளர்க்க ஒவ்வொரு கணுப் பிளவிலும் தகவல் ஆதாயத்தை அதிகரிக்க சிறந்த கேள்வியைத் தேர்ந்தெடுப்பதற்கான நடைமுறையையும் கொண்டுள்ளது. இந்த தகவல் என்ட்ரோபி என அளவிடப்பட வேண்டும் (சமன்பாடு 1). தகவல் அதிகரிக்கும் வரையும் - என்ட்ரோபியில் குறையும் வரையும் ஒரு தொடக்க நிலையிலிருந்து விழும் வரைபிளவு தொடர்கிறது (சமன்பாடு 2).

தீர்மானக்களுக்கான கேள்விகள் (Questions for decision trees)

எழுத்திலிருந்து ஒலி மாற்றத்திற்கான அடிப்படை ஆம்-இல்லை கேள்வி பின்வருமாறு அமையும்: "முதல் இடது எழுத்து 'm' என்பதா?". தமிழில் எழுத்துக்களுக்கான கேள்விகள் ஒலிகளுக்கான கேள்விகளை மிகவும் ஒத்திருப்பதால் அவை சேர்க்கப்படவில்லை. கேள்வி நிலைகளின் வீச்சு நீண்ட தூர ஒலியனியல் மாறுபாடுகளை உட்படுத்தப் போதுமானதாக இருக்க வேண்டும். 5 எழுத்து சாளரம் (இடது சூழலுக்கு 2 மற்றும் வலது சூழலுக்கு 2) பொதுவாக போதுமானது என்று கண்டறியப்பட்டுள்ளது. ஒரு பழமையான தொகுப்பு என்பது அனைத்து சிங்கிள்டன் கேள்விகளின் (ஒற்றை எழுத்து/நெடுங்கணக்கு பற்றிய கேள்வி) கணமக/தொகுப்பாக (set) இருக்கும். எளிமையான கேள்விகளை மட்டுமே அனுமதிப்பதில் ஒரு சிக்கல் இதன் விளைவாக வெவ்வேறு இடங்களில் ஒரே இலைகள் விளையும் படி தரவு அதிகமாகத் துண்டு படுத்தப்படலாம். வெவ்வேறு கணுக்களில் தரவைப் பிரிப்பது மிகை/தேவையற்ற கொத்துகளுக்கும் குறைக்கப்பட்ட பயிற்சிக்கும் குறைந்த துல்லியம் என்ட்ரோபி குறைப்புக்கும் வழிவகுக்கும்.

ஒலிப்புடன் தொடர்புடைய கேள்விகளுடன் "இடது எழுத்து ஒரு மூக்கொலியா?" போன்ற சிக்கலான கேள்விக் கணுவை அனுமதித்தால் கிளையமைப்பின் ஆழம் பெரிதும் குறைக்கப்பட்டு

செயல்திறன் மேம்படுத்தப்படும் என்பது உற்றுநோக்கப்பட்டது. சிக்கலான கேள்விகள் தீர்மான கிளையமைப்பு வழிமுறை பேராசைத் தன்மையால் ஏற்படும் தரவைத் துண்டுபடுத்தும் சிக்கலைத் தணிக்கும். உயிரொலிகள், மெய்யொலிகள், மூக்கொலொகள், அடைப்பொலிகள், உடன்படுத்திகள் மற்றும் வளைநா ஒலிகள் ஆகியவை ஒலிக் கொத்துகளின் சில கலப்பு கேள்விகளை உருவாக்க பயன்படுகிறது.

ஒலி வரிசைகளின் உருவாக்கம் Generation of phone sequences

தீர்மானக் கிளையமைப்புகள் பயிற்சி பெற்ற பிறகு, சொற்றொகையில் உள்ள சொற்களுக்கான ஒலி வரிசைகளை உருவாக்குவதற்கு அவற்றைப் பயன்படுத்தலாம். ஒவ்வொரு வார்த்தைக்கும், ஒரே நேரத்தி சொல்லின் இடமிருந்து வலமாகச் செல்வதன் மூலம் ஒரு ஒலியன் வரிசை உருவாக்கப்படுகிறது. சூழல் தகவல் அடிப்படையில் வேர்க் கணுவிலிருந்து தொடங்கி கிளையமைப்பு கடந்துசெல்லப்படுகிறது. ஒரு இலையை (இறுதிநிலை) அடைகிறது வரை நிரல் (program) கிளையமைப்பு வழியாகச் செல்கிறது; இலையை அடையாளம் காணுதல் தற்போதைய எழுத்தின் ஒலிபெயர்ப்பாகும். பின்னர் செயல்முறை இதேபோன்று அடுத்த எழுத்துக்கு நகர்கிறது. ஒரு வேளை போலி ஒலியனாக இருந்தால் இறுதி ஒலிபெயர்ப்பை அடைய மாற்று உச்சரிப்புகள் உருவாக்கப்படுகின்றது. எ.கா. சென்னை /ch*s ae n h ax y / என்பது /ch ae n h ax y / மற்றும் / s ae n h ax y / ஆக விரிவாக்கப்படுகிறது.

சோதனைகள்

சோதனைகளுக்கு EMILLE விரிதரவு/தரவுத்தொகுதி (corpus) பயன்படுத்தப்பட்டது. இந்த தரவுத்தொகுதி யூனிகோடால் உருப்படுத்தம் செய்யப்பட்டுள்ளது மற்றும் மென்பொருளால் எளிதாகக் கையாளப்படுவதற்காக ஆஸ்கி போன்ற உருப்படுத்ததிற்கு மாற்றப்பட்டது. தரவுத்தொகுதியிலிருந்து 85,000 சொற்கள் தேர்ந்தெடுக்கப்பட்டது மற்றும் தீர்மானக் கிளையமைப்புக்கு பயிற்சி கொடுக்க அகராதி உருவாக்கப்பட்டது. தீர்மானக் கிளையமைப்புகள் நியூசிலாந்திலுள்ள வைகாடோ பல்கலைக்கழகத்தால் (University of Waikato, New Zealand) உருவாக்கப்பட்ட இயந்திர கற்றல் வழிமுறைகளுடன் சோதனைக்கான கருவியான வெக்காவைப் (Weka) பயன்படுத்தி செயல்படுத்தப்பட்டது.

விதி அடிப்படையிலான எழுத்திலிருந்து ஒலியன் மாற்ற ஒழுங்கமைப்பு

எளிய எழுத்து-ஒலியன் சார்புகளைக் கண்டுபிடிக்க மொழியியல் விதிகள் அடிப்படையில் ஒன்று வலது ஒன்று இடது சூழலைப் பயன்படுத்தி விதி அடிப்படையிலான அமைப்பு தயாரிக்கப்பட்டது. ஒரு எழுத்து இடைவெளியைத் தாண்டி விதிகளைக் கண்டுபிடிப்பது விதி

அடிப்படையிலான ஒழுங்கமைப்பில் மிகவும் கடினம். ஒழுங்குமுறை அதன் துல்லியதிற்குப் பரிசோதிக்கப்பட்டு அட்டவணை 2 இல் பட்டியலிடப்பட்டுள்ளது.

தீர்மானக் கிளையமைப்பு அடிப்படையிலான எழுத்திலிருந்து ஒலியன் மாற்ற ஒழுங்கமைப்பு

ஒலிபெயர்ப்பைத் உருவாக்குவதற்காகக் கைமுறையாக உருவாக்கப்பட்ட அகராதியைப் பயன்படுத்தி தீர்மானக் கிளையமைப்புகளுக்குப் பயிற்சி அளிக்கப்படுகின்றன. ஆரம்பத்தில் விதி அடிப்படையிலான அமைப்பைப் பயன்படுத்திப் பூட்ஸ்ட்ராப் செய்யப்பட்டது மற்றும் ஒரு வல்லுன மொழியியலாளரால் கைமுறையாகச் சீர்படுத்தப்பட்டது. முதலில், தீர்மானக் கிளைகளை உருவாக்க எழுத்து அடையாளத்தைப் பற்றிய சிங்கிள்டன் கேள்விகள் பயன்படுத்தப்பட்டது. ஒழுங்கமைப்பு பரிசோதிக்கப்பட்டது மற்றும் துல்லியம் விதி அடிப்படையிலான அணுகுமுறையுடன் ஒப்பிடப்பட்டது.

மேலும், துல்லியத்தை அதிகரிக்க, மூக்கொலிகள் முன் உயிரொலிகள், பின் உயிரொலிகள், மெய்யொலிகள், அடைப்பொலிகள் போன்ற ஒலியன்ககளின் ஒலிப்பு ஊக்கமளிக்கும் கொத்துகள் உருவாகப்பட்டது மற்றும் அவை கிளையமைப்புகளுக்குப் பயிற்சியளிக்கக் கேள்விகளாகப் (சிக்கலான கேள்விகள்) பயன்படுத்தப்பட்டன. அட்டவணை 2 இந்த சோதனைக்கான செயல்திறன் முடிவுகளை பட்டியலிடுகிறது.

எழுத்திலிருந்து ஒலியன் மாற்ற ஒழுங்கமைப்பு	சொல் நிலை	ஒலி நிலை
விதி அடிப்படையிலானது	94%	95.5%
தீர்மானக்கிளையமைப்பு சிங்கிள்டன் கேளிவிகள்	96.5%	98%
தீர்மானக்கிளையமைப்பு கலவைத்தன்மையான கேள்விகள்	98%	98.5%

3.5.8.2.4. அருண்குமாரின் தமிழ் உரை-பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறை

அருண்குமார் (2010) 'தமிழ் உரை-பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறை' (Text to speeh for Tamil)' என்ற அவரது ஆய்வேட்டில் ஆய்வில் தமிழ் உரை-பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறை பற்றி விரிவாகப் பேசுகின்றார். தற்போதைய உரை-பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறை ஆய்வு இதற்கு முன் செய்யப்பட்ட உரை-பேச்சு ஒழுங்குமுறைகளைக் கருத்தில் கொண்டு அவற்றைவிட மேம்பட்ட ஒழுங்குமுறையை உருவாக்கும் எண்ணத்தில் திட்டமிடப்பட்டு

நடைமுறை படுத்தப்பட்டுள்ளது என்று அவர் கூறுகின்றார். இது கணிபொறி திரையில் படிப்பதைவிடக் கேட்பதை விரும்புபவர்களுக்கும் டிஜிட்டல் நூல்களைப் படிக்க விரும்பாமல் கேட்க விரும்புபவர்களுக்கும் கணிபொறி திரையில் படிக்க இயலாக் குறைபாடு உடையவர்களுக்கும் உதவும் நோக்கத்துடன் மேற்கொள்ளப்பட்டது. இது குறித்து இப்பகுதியில் விளக்கப்பட்டுள்ளது.

ஒழுங்குமுறைத் திட்டவரைவு (system design)

ஒழுங்குமுறை திட்டவரைவு ஒழுங்குமுறையின் மேலோட்டப் பார்வை, ஒழுங்குமுறை கட்டமைப்பு இவற்றை உட்படுத்தும்

ஒழுங்குமுறையின் மேலோட்டப் பார்வை

இந்த ஒழுங்குமுறை தட்டச்சு செய்யப்பட்ட தமிழ் சொற்களையோ புரிந்துகொள்ளும் மற்றும் அதற்குப் பொருத்தமான பேச்சை உருவாகும். இவ்வொழுங்குமுறையின் அமைப்பு பின்னர் தரப்பட்டுள்ளது. இந்த ஒழுங்குமுறை இரண்டு நிலைகளைக் கொண்டது.

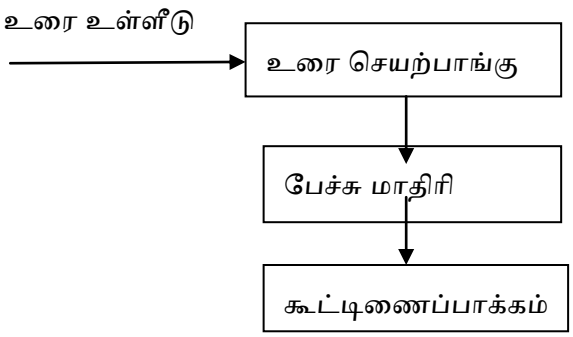
- உரைப் பகுப்பாய்வு நிலை (Text processing)
- இணைப்பாக்க நிலை (Concatenative phase)

கீழே தரப்பட்ட ஒழுங்குமுறைக்கு உள்ளீடு நேரடியாகத் தர இயலாது; சில உரை பகுப்பாய்வு செய்யப்படவேண்டும். பின்னர் இந்தப் பகுப்பாய்வு செய்யப்பட்ட உள்ளிடப்பட்ட உரை பேச்சு செயற்பாங்கிற்கு உள்ளீடாகத் தரப்படும். செயற்பாங்கு முழுவதையும் முடித்த பின்னர் செயற்பாங்கு செய்யப்பட்ட உள்ளீடு உரை அடிப்படையில் தரவு அடித்தளத்திலிருந்து பொருத்தமான பேச்சுக் கோப்பு எடுக்கப்படவேண்டும். ஒவ்வொரு ஒலியன் அல்லது ஒலியிருமையும் ஒரு தனி பேச்சு கோப்பு தரும்படி இணைக்கப்பட வேண்டும். பேச்சுக் கோப்பின் இந்த ஒன்றிணைப்புச் செயற்பாங்கு ஒன்றிணைப்பு நிலையில் செய்யப்படுகின்றது.

ஒழுங்குமுறைக் கட்டமைப்பு (system architecture)

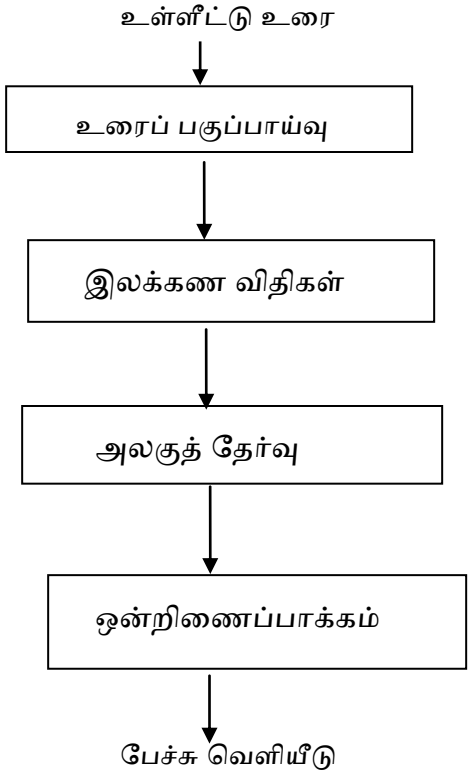
உரையின் செயற்பாங்கு பெர்ல் வழிமுறை திட்டத்தைப் பயன்படுத்தி (Per program) நிறைவேற்றப்பட்டுள்ளது. ஒவ்வொரு செயற்பாங்கிற்கும் அடிப்படையான நடைமுறை ஒழுங்குமுறைக்கு ஒருகட்டமைப்பை உருவாக்குவதாகும்; ஏனென்றால் அது ஒரு கட்டடத்தின் வரைபடம் போன்றது. உரை-பேச்சு ஒழுங்குமுறையின் கட்டமைப்பு படம் 1-இலும் 2-இலும் காட்டப்பட்டுள்ளது.

படம் 1: உரை-பேச்சு ஒழுங்கமைப்பின் அடிப்படைஅமைப்பு



இந்தப் பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறை பலவகைப் பயன்பாடுகளைக் கொண்டது. பேச்சு கூட்டிணைப்பாக்கிகள் பேச்சுப் புரிவான்களிலும் தொலைபேசி அடிப்படையிலான மக்களுடன் கருத்தாடல் செய்யும் கருத்தாடல் செயலர்களிடையேயும் பயன்படுத்தப்படுகின்றது. பேச்சு கூட்டிணைப்பாக்கிகள் மக்களிடம் பேசும் கருத்தாடல் அல்லாத பயன்பாடுகளான பார்வை குறைபாடுள்ளவர்களுக்கு உரக்கப் படித்தல், வீடியோ/காட்சி விளையாட்டுகள் அல்லது குழந்தைகளின் பொம்மைகள் இவற்றில் முக்கியத்துவம் வாய்ந்தது. இறுதியாகப் பேச்சு கூட்டிணைப்பாக்கி நரம்பு கோளாறு காரணமாகப் படிக்க இயலாதவர்களுக்கு வேண்டி பேச இயலும்.

படம் 2: உரை-பேச்சு ஒழுங்கமைப்பில் உள்ள செயற்பாங்குகள்



பேச்சுக் கூட்டிணைப்பாக்கச் செயற்பாங்கு

உரை-பேச்சு ஒழுங்கமைப்பின் முக்கியமான செயற்பாங்கு உள்ளீட்டு உரையைப் பேச்சுக் கோப்புகளாக மாற்றுவதாகும். இந்த ஒழுங்குமுறையில் தமிழ் உரையை தமிழ்ப் பேச்சாக மாற்றும் செயற்பாங்கு நிறைவேற்றப்படுகின்றது. இதை கீழே தரப்பட்டுள்ள எடுத்துக்காட்டிலிருந்து புரிந்துகொள்ள இயலும். பேச்சுக் கூட்டிணைப்பாக்கியின் செயற்பாடு பின்வருமாறு உரையை அலைவடிவில் மாற்றுவதாகும்.

தமிழ் மொழி இனிமையானது.

பேச்சுக் கூட்டிணைப்பாக்க ஒழுங்குமுறை இரண்டு நடவடிக்கைகளை உள்ளடக்கும்:

1. முதலில் உள்ளீட்டு உரையை ஒலியனியல் அக உருப்படுத்தத்தமாக (phonemic internal representation) மாற்றுவது.
2. பின்னர் இந்த அக உருப்படுத்ததை அலைவடிவமாக மாற்றுவது.

முதல் நடவடிக்கை உரை ஆய்வு (text analysis) என்றும் இரண்டாம் நடவடிக்கை அலைவடிவ கூட்டிணைப்பாக்கம் (wave synthesis) என்றும் அழைக்கப்படும்.

உரைச் செயற்பாங்கு (text processing)

ஒலியனியல் அக உருப்படுத்தத்தை உருவாக்க வேண்டி பல வழிகளில் பக்குவப்படுத்தப்படாத உரை செயற்பாங்கிற்கு உட்படுத்தப்பட வேண்டும் அல்லது இயல்பாக்கப்படவேண்டும். உரை இயல்பாக்கத்தில் முதல் செயல்பாடு வாக்கியத்தை டோக்கன்களாகப்/கூறுகளாகப் பிரித்தல். கூட்டிணைப்பக்கத்திற்கு வேண்டி பத்தியை தனி கூற்றுகளாகப் பிரிப்பதற்கு வாக்கியம் முழுப்புள்ளிக்குப் பிறகு முடிகின்றது என்று அறியவேண்டும்; ரூ. என்ற கூற்றில் வரும் ரூ-வுக்குப் பிறகு வரும் புள்ளிக்குப் பிறகல்ல. ஒரு முழுபுள்ளிக்கு மாறாக காற்புள்ளி நிறுத்தற்குறியாக வந்தாலும் சேகரிக்கப்பட்ட சொற்களின் இறுதியில் வாக்கியம் முடிகின்றது என்றும் அறிய வேண்டும். இரண்டாவது இயல்பாக்கச் செயற்பாடு நிலைபேறற்ற சொற்களைக் கையாளுதல் ஆகும். நிலைபேறற்ற சொற்கள் எண்கள், தலைப்பெழுத்தன்கள், சுருக்கச்சொற்கள் போன்றனவற்றை உட்படுத்தும். எடுத்துக்காட்டாக 'மார்ச் 31' என்பது 'மார்ச் மூன்று ஒன்று' என்றல்லாமல் மார்ச் முப்பதி ஒன்று என்று மாற்றப்படவேண்டும். 'ரூ. 1' என்ற சுருக்கம் 'ரூ ஒன்று' என்றல்லாமல் 'ரூபாய் ஒன்று' என்று மாற்றப்படவேண்டும். உரை இயல்பாக்கத்தில் இரண்டாவது நடவடிக்கை நிலைபேறல்லாத சொற்களை இயல்பாக்குதல் ஆகும். எண்கள் அல்லது சுருக்கச் சொற்கள் அவை பேச்சாக மாற்றப்படும் முன்னர் விரிவாக்கப்படவேண்டும்.

எனவே நாம் தமிழின் தனி எழுத்துக்களை ஒலிப்பதிவு செய்வதற்குப் பதிலாக எழுத்துக்களின் அதாவது ஒலியிருமைகளின் ஒன்றிணைப்புகளைப் பதிவுசெய்கின்றோம். இவ்வாறு செய்வது ஒழுங்குமுறையின் இயல்புத் தன்மையை மேம்படுத்தும்.

இலக்கண விதிகள்

உரைச் செயற்பாங்கிற்குப் பின்னர் நமக்கு செயற்பாங்கிற்கு உட்படுத்தப்பட்ட தமிழ் உரை கிடைக்கப்பெறும். அதை நேரடியாக மாற்றச் செயற்பாங்கிற்குத் தர இயலாது. தமிழ் உரை ரோமன் வடிவுக்கு மாற்றப்படவேண்டும். ஏனென்றால் கணிப்பொறி தமிழ் தரவை அப்படியே செயற்பாங்கு செய்ய இயலாது; எனவே மாற்றம் கட்டாயம் செய்யப்படவேண்டும். தமிழில் எழுத்துக்கள் வெவ்வேறு இடங்களில் வேவேறுவிதமாக உச்சரிக்கப்படுகின்றன.

மயக்கம் அடிப்படையில் தமிழ் ஒழுங்குமுறையை அமைப்பது ஆங்கில ஒழுங்குமுறையை அமைப்பதைக் காட்டிலும் எளியதாகும். ஆங்கில எழுத்துமுறை நெடுங்கணக்கு

அடிப்படையிலானது; எழுத்துக்கும் உச்சரிக்கும் நேரடித் தொடர்பு இல்லை; பல எழுத்துக்களுக்கு உச்சரிப்பு இல்லை. எடுத்துக்காட்டாக ஆங்கிலத்தில் thought என்ற சொல்லை எடுத்துக்கொண்டால் சில எழுத்துக்களுக்கு நேரடி உச்சரிப்பு இல்லாததை அறிய இயலும். தமிழ் எழுத்துக்கள் அசை அடிப்படையிலானவை; எழுத்துக்களுக்கு நேரடி உச்சரிப்பு இருக்கின்றது. தமிழில் உள்ள எல்லா சொற்களையும் உச்சரிக்கத் தமிழுக்குக் குறைவான எழுத்துக்களே தேவைப்படும்; ஆனால் எழுத்துக்களின் உச்சரிப்பு அவை வரும் இடம் அடிப்படையில் மாறும். இதை ஒலியனியல் விதிகளால் அல்லது இலக்கண விதிகளால் கையாளலாம். இந்த விதிகளைக் கொண்டு நம்மால் வரும் இடம் அடிப்படையில் தமிழ் எழுத்துக்களின் உச்சரிப்பை ஊகிக்க இயலும். இந்த செயற்பாங்கு இலக்கணத் தொகுதியில் செய்யப்படும்.

பேச்சு தரவு அடிப்படைத்தளம்

தமிழில் உள்ள எல்லா எழுதுக்களும் ஒலிப்பதிவு செய்யப்பட்டு தனியே ஒரு தரவு அடிப்படைத் தளத்தில் வைக்கப்படவேண்டும். உரை/பனுவல் செயற்பாங்கு, இலக்கண மாற்றம், ரோமன் எழுத்துக்கு மாற்றம் இவற்றிற்குப் பிறகு ஒழுங்குமுறை பேச்சு தரவு அடித்தளத்திலிருந்து பொருத்தமுள்ள பேச்சுக் கோப்பை எடுக்கவேண்டும். தமிழ் எழுத்துக்களின் பேச்சு ஒலிகளின் ஒலிப்பதிவு கூடுதல் குறிகை செயற்பாங்கிற்குவேண்டி சில தேவைகளை அவசியப்படும்.

ஒலிப்பதிப்பு ஒலிப் பாதுகாப்பு அறையில் செய்யப்படவேண்டும்; தமிழ் எழுத்துக்களின் உச்சரிப்பு சிறப்பாக இருக்கவேண்டும்; பேச்சுக் கோப்பில் இரைச்சலைத் தவிர்க்க நல்ல ஒலிவாங்கி பயன்பத்தப்படவேண்டும். பேச்சுக் கூட்டிணைப்பாகத்தில் துல்லியம் பேச்சு வெளியீட்டைன் இயற்கைத் தன்மை அடிப்படையில் கணிக்கப்படுகின்றது. எனவே தமிழின் தனி எழுத்துக்களை ஒலிப்பதிவு செயவதற்குப் பதிலாக நாம் எழுத்துக்களின் ஒன்றிணைப்பை அதாவது ஒலியிருமையை ஒலிப்பதிவு செய்கின்றோம். இது உரை-பேச்சு ஒழுங்கமைப்பின் தரத்தை உயர்த்தும்.

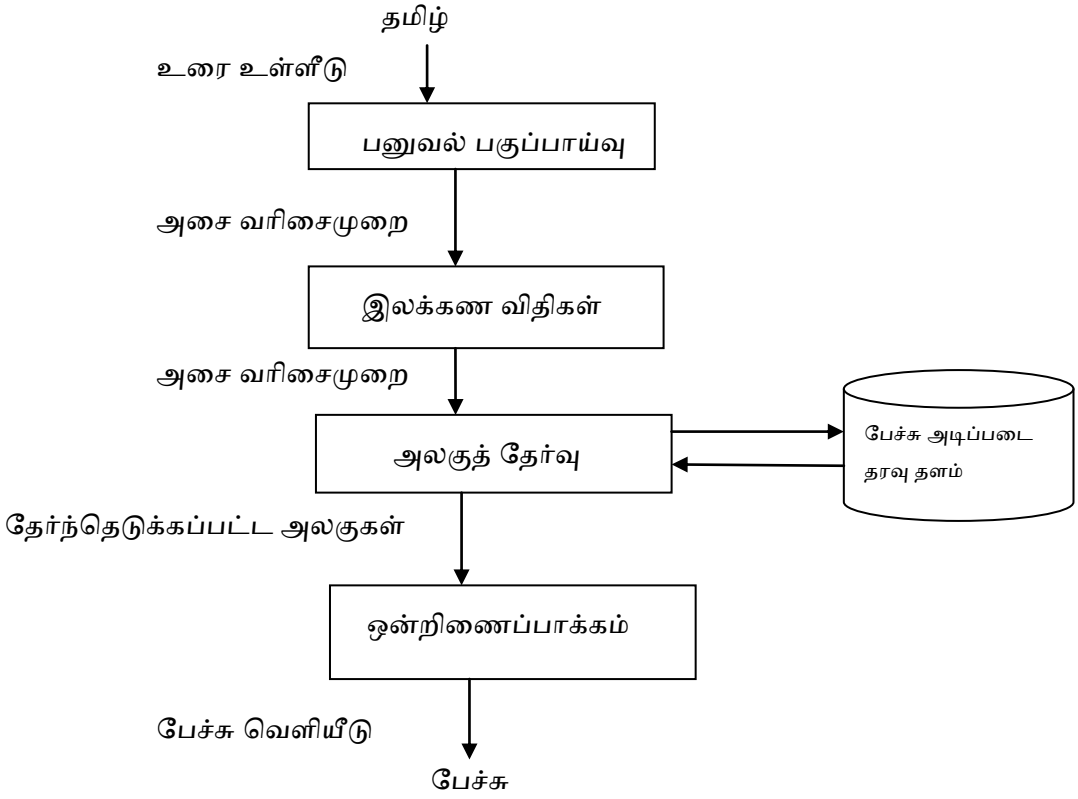
ஒன்றிணைப்பாக்கம்

உரை-பேச்சு ஒழுங்குமுறையின் இறுதி நிலை ஒன்றிணைப்பாக்கம் ஆகும். இது இயல்பான வடிவைப் பெற்ற பின் தரவு அடித்தளத்திலிருந்து பொருத்தமான பேச்சு கோப்பைப் தெரிந்தெடுப்பதையும் இந்த எல்லா பேச்சுக் கோப்புகளையும் ஒரே கோப்பாக ஒன்றிணைப்பதையும் உட்படுத்தும். ஒன்றிணைப்பக்கச் செயற்பாங்கு மேட்லாபின் வழிமுறை திட்டத்தின் உதவியால் நடைமுறை படுத்தப்படும். இரண்டு பேச்சுக் கோப்புகளை ஒன்றிணைக்கும் போது ஏற்படும் தடுமாற்றத்தைத் தவிர்க்க மேட்லப் பயன்படுத்தப்படுகின்றது.

நடைமுறைப்படுத்தல்

தமிழில் உரைலிருந்து பேச்சு கூட்டிணைப்பு உருவாக்கத்தை ஒன்றிணைப்பு நெறிமுறையைப் பயன்படுத்தி நடைமுறைப் படுத்தும் முறை பின்வரும் படத்தில் காட்டப்பட்டுள்ளது.

உரை-பேச்சு ஒழுங்குமுறையின் தொகுதிப் படம்



உரைப் பகுப்பாய்வு

உள்ளீடு செய்யப்பட்ட தமிழை அலகுத் தேர்வுக்கு நேரடியாகத் தர இயலாது; பதப்படுத்தப்படாத உரை/பனுவல் செயல்முறை படுத்தப்படவேண்டும் மற்றும் அது மேலும் செயல்முறைப் படுத்தலுக்குத் தரப்படும். உரையிலிருந்து/பனுவலிலிருந்து பேச்சு ஒழுங்கமைப்புக்கு நாம் உள்ளீட்டை அசைகளாக அவசியப்படுகின்றோம். சொல் அசைகளின் வரிசையாகத் தரப்படும் போது இந்த ஒழுங்கமைப்பு எங்கே பிரிக்கவேண்டும் என்று அறியாது; எனவே இந்த வரிசை அசையாக்கத்தை செய்ய நாம் இயந்திரத்திற்கு சொல்லையோ வாக்கியத்தையோ எங்கே பிரிக்கவேண்டும் என்று சொல்ல வேண்டும்.

முன்செயல்முறைப் படுத்தம்

அசைபிரிப்பு செயல்முறை படுத்தத்திற்குத் தருவதற்கும் முன் சில முன்செயல்முறைப்படுத்தம் செய்யப்படவேண்டும். அதாவது சொற் பிரிப்பு செய்யப்படவேண்டும். பனுவல் இயல்பாக்கத்திற்கு முதல் செயல்பாடு வாக்கியத்தைச் சொற்களாகப் பிரிப்பதாகும். கூட்டிணைப்பு உருவாக்கத்திற்கு வேண்டி தமிழ் பத்திகளைத் தனித்தனிக் கூற்றுகளாகப் பிரிப்பதற்கு முதல் வாக்கியம் முற்றுப் புள்ளிக்குப் பின்னர் முடிவடையும் என்றும் முற்றுப்புள்ளி அல்லாத தா.நா... போன்றவற்றில் உள்ள புள்ளிகளில் முடிவுறாது எனவும் தெரிந்துகொள்வது அவசியம். சொல்லைப் பிரித்தெடுப்பது ஓரவுக்கு எளிது; அதாவது பெரும்பாலான வாக்கியங்கள் முற்றுப் புள்ளியில் முடிவுறும். இருப்பினும் இது சுருக்கக் குறியீடுகளாலும் அரைப்புள்ளிகளாலும் அல்லது முந்தைய நேர்வைப் போன்று பிற நிறுத்தற் குறிகளாலும் முடிவுறும் சில நேர்வுகள் உள்ளன.

சுருக்கக்குறியீடுகளை விரிவாக்கியும் தேவையற்ற நிறுத்தற் குறிகளை நீக்கியும் இச்சிக்கலுக்குத் தீர்வுகாண இயலும். எடுத்துக்காட்டாக பின்வரும் வாக்கியத்தை எடுத்துக்கொள்வோம்.

அவன் வீடு சென்றான்.

இவ்வாக்கியத்தை முழுப்புள்ளி/முற்றுப்புள்ளியுடன் தொடர்புபடுத்திச் சொற்களைத் தனித்தனியாகப் பிரிக்க இயலும்; எனவே இவ்வாக்கியத்தில் சொற்களைப் தனித்தனியாகப் பிரிப்பதில் சிக்கல் இல்லை. ஆனால் பின்வரும் வாக்கியம் சிக்கல் உள்ளதாகும்.

அவன் தா.நா. சென்றான்.

இவ்வாக்கியத்தில் சுருக்கக் குறியீடுகள் வருகின்றன; அவற்றின் பின் முழுப்புள்ளி வருகின்றன. சுருக்கக்குறியீடுகளை விரிவாக்கி இச்சிக்கலைத் தீர்க்கலாம். இதனால் இதை முந்தைய நேர்வைப் போன்று சொற்களாகப் பிரிக்க இயலும். எல்லா தமிழ் சுருக்கக் குறியீடுகளையும் விரிவாக்க இயலாது; ஏனென்றால் சில அதிகமாகப் பயன்படுத்தப்படும் சுருக்கக்குறியீடுகள் தனி தரவு அடித்தளத்தில் சேகரிக்கபட்டிருக்கும். அந்த சுருக்கக் குறியீடு இருந்தால் ஒழுங்கமைப்பு பனுவலை இடம்பெயர்க்கும்; இல்லாவிடில் அது மூலப் பனுவலை அவ்வாறே விட்டுவிடும். எல்லா சுருக்கக்குறியீடுகளையும் தரவு அடித்தளத்தில் சேர்ப்பது கடினம்; எனவே சில அதிகமாகப் பயன்படுத்தப்படும் சுருக்கக் குறியீடுகள் பயன்படுத்தப்படுகின்றன. பின்வருவன எடுத்துகாட்டுகளாகும்.

தா.நா. ⟶ தமிழ் நாடு

ரூ. ⟶ ரூபாய்

குழப்பத்தைத் தவிர்க்கவும் பேச்சில் இயற்கைத்தன்மை கெடாதிருக்கவும் தரப்பட்டுள்ள தமிழ் பத்தியிலிருந்து தேவையற்ற நிறுத்தற் குறியீடுகள் (:, ;, “, ‘ போன்றன) நீக்கப்படவேண்டும். உள்ளீட்டின் ஒவ்வொரு பனுவலும் ஒன்றிணைப்புக்கு வேண்டி எதாவது ஒலி கோப்புக்கு ஒதுக்கப்படவேண்டும். நாம் இந்த நிறுத்தற்குறிகளை வைத்திருந்து அவற்றிற்கு ஏதாவது பேச்சுப் பனுவல் ஒதுக்கினால் அது இறுதி வெளியீட்டுப் பேசுக் கோப்பின் இயல்பின்மைக்குக் கொண்டுசெல்லும். தரப்பட்ட வழிமுறைவரைவு இவற்றையெல்லாம் செய்யும் மற்றும் அது உள்ளீட்டுப் பனுவலில் உள்ள இடைவெளிகளுக்கு @ குறியீட்டை ஒதுக்கும். சில பத்திகளில் கூடுதல் இடைவெளி இருக்கும் மற்றும் அதிகப்படியான நிறுத்தற்குறிகள் இருக்கும்; நிறுத்தற்குறி வெற்று இடைவெளியாக ஆக்கப்படும் மற்றும் ஒரு தனி இடைவெளியாகக் கூடுதல் இடைவெளிகள் ஆக்கப்படும். இறுதியாக இடைவெளிகளுக்கு @ குறியீடு ஒதுக்கப்படும்; இந்த @ குறியீடு மாட்லாபில் மௌனமாக ஒதுக்கப்படும். இது மேலும் வரும் செயல்முறையில் உதவியாகா இருக்கும்.

பனுவல் வழமையாக்குவதில் அடுத்த நடவடிக்கை தரமற்ற சொற்களை வழமையாக்குவதாகும். எண்கள் அல்லது சுருக்கக் குறியீடுகள் போன்று தரமற்ற சொற்கள் டோக்கன்களாகும்; அவைகள் உச்சரிக்கப்படுவதற்கு முன்னர் தமிழ்ச் சொற்களின் வரிசையாக விரிவாக்கப்பட வேண்டும். இந்த தரமற்ற சொற்களைப் பற்றிய சிக்கல் என்னவென்றால் அவை பெரும்பான்மையும் மயக்கம் உள்ளவை. எடுத்துக்காட்டாக 1750 என்ற எண் சூழல் அடிப்படையில் குறைந்தது மூன்று வேறுபட்ட வழிகளில் பேசப்பட இயலும்.

பதினேழு ஐம்பது.

ஆயிரத்து எழுநூற்றி ஐம்பது.

மூன்று ஏழு ஐந்து பூஜியம்.

இவ்வெல்லா நேர்வுகளுக்கும் ஒரு தனியான நெறிமுறையைக் கருத்திற்கொண்டு இச்சிக்கலைக் கையாளலாம். தமிழுக்கு எண் ஒழுங்கமைப்பு பைத்தான் குறியனில் ஏற்கனவே உருவாக்கப்பட்டுள்ளது. இது நூற்றுக்கு நூறு துல்லியத்தைத் தரும்; இதைப் பனுவலிலிருந்து பேச்சு ஒழுங்குமுறையில் உட்படுத்த இயலும்; அனால் இது ஒழுங்குமுறையைத் தாமதப்படுத்தும்.

அசை பிரித்தல் (Syllabification)

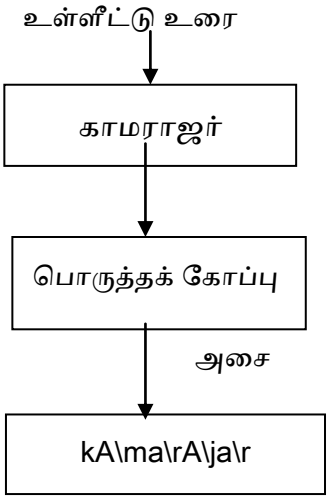
இதைப் பல வழிகளில் செய்ய இயலும்; எடுத்துக்காட்டாக அசைப்பிரித்தல் வழிமுறைவரைவு ஒரு சொல்லைக் குறைந்த எண்ணிக்கையிலான பிரிப்புகள் வரும்படி சொல்லைப் பிரிக்கும்; ஏனென்றால் குறைந்த எண்ணிக்கையிலான ஒன்றுசேர்ப்புகள் குறைந்த

செயற்கை விளைவுகளைக் கொண்டிருக்கும். வழிமுறைவரைவு சொல்லை உருவாக்கும் பல்லசை அலகுகளுக்கு வேண்டி தேடும்; அலகுகளின் கண்டுபிடிப்புக்கு வேண்டி தரவு அடித்தளத்தை மீண்டும் சரிபார்க்கும்; பின்னர் சொல்லை அதன்படி பிரிக்கும். பல்லசை அலகுகள் கிடைக்காவிடில் வழிமுறைவரைவு இயல்பாகச் சிறிய அலகுகளை எடுக்கும். இதன் அர்த்தம் அசை அலகுகளுடன் மொழியின் எல்லா ஒலிகளாலும் தரவு அடித்தளம் நிறைக்கப்பட்டால் வழிமுறைவரைவு பெரிய அலகுகள் கிடைக்காவிடில் ஒலிகளைத் தெரிவுசெய்யும்.

ஆனால் தற்போதைய ஒழுங்கமைப்பில் இந்த நெறிமுறை பின்பற்றப்படவில்லை; மாறாக தமிழ் எழுத்துக்களைக் கொண்டிருக்கும் பொருத்தக் கோப்பு (mapping file) பயன்படுத்தப்படுகின்றது. உள்ளீடு செய்யப்பட்ட தமிழ் எழுத்துக்களுக்குப் பதில் தொடர்புள்ள உரோமன் எழுத்துக்கள் எடுக்கப்பட்டு அசைகளாக மாற்றப்படுகின்றன. இந்தப் பொருத்தக் கோப்பில் முக்கிய செய்தி எழுத்துக்களின் வரிசை தமிழ் எழுத்துக்களின் நீளம் அடிப்படையில் அமைய வேண்டும். எழுத்துக்கள் தேவைப்படும் முறையில் வரிசைப்படுத்தப்படாவிடில் ஒழுங்கமைப்புத் தவறு நிகழும் மற்றும் வெளியீடு கிடைக்கப்பெறாது.

எடுத்துக்காட்டாக உள்ளீடாக *அம்மா* என்ற தமிழ்ச் சொல் இருந்தால் *அ* மற்றும் *ம்* என்பன சிக்கல் வாய்ந்தன அல்ல. *ம* என்பது *மா* முன் வந்தால் பெரிய சிக்கல் வரும். வழிமுறைவரைவு *ம*-க்கு வேண்டி தேடும்; பின்னர் *மா*-ஐ *ம*-ஆக இடம் பெயர்க்கும். இதனால் ஒரு புரிந்துகொள்ளப்படாத எழுத்து வரும்; எனவே ஒழுங்கமைப்பு அலகுத் தேர்வுக்கு வேண்டி இதைச் செயற்பாங்கு செய்ய இயலாது; எனவேதான் எழுத்துக்களின் வரிசை முறை கட்டாயம் ஆகும்.

காமராஜர் என்ற சொல்லின் அசை பிரிப்பு



Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

தமிழ் எழுத்து-உரோமன் எழுத்துப் பொருத்தம்

தமிழ் எழுத்து	உரோமன் எழுத்து
இ	i
ஈ	I
உ	u
ஊ	U
எ	e
ஏ	E
ஐ	ai
ஒ	o
ஓ	O
ஔ	au
ஃ	q
க	ka
ங	nga
ச	sa
ஞ	nja
ட	da
ண	Na
த	tha
ந	wa
ப	pa
ம	ma
ய	ya
ர	ra
ல	la
சா	sA
சி	si

சீ	sI
சு	su
சூ	sU
செ	se
சே	sE
சை	sai
சொ	so
சோ	sO
சௌ	sau

தொடர்புள்ள உரோமன் எழுத்துடன் பொருத்தப்பட்டுள்ள தமிழ் எழுத்து பொருத்தக் கோப்பின் பாகமாகும். இதில் எழுத்துக்கள் எல்லாம் நீளம் அடிப்படையில் வரிசைப்படுத்தப்பட்டுள்ளன.

இலக்கண விதிகள்

தமிழில் பனுவல் வடிவமைப்பில் உள்ள ஒரே எழுத்து அவை வரும் வேறுபட்ட இடங்கள் அடிப்படையில் வேறுபட்ட உச்சரிப்புகளைக் கொண்டிருக்கும். இதைப் பல வழிகளில் தீர்வு காணலாம்: விதிகளை எழுதியோ அல்லது எஸ்.வி.எம். வழிமுறைவரைவு, எச்.எம்.எம் போன்ற இயந்திரம் கற்றல் அணுகுமுறைகளில் ஒன்றாலோ தீர்வுகணாலாம். இயந்திரம் கற்றல் அணுகுமுறையைப் பயன்படுத்தும் முந்தைய பல அணுகுமுறைகள் போலல்லாமல் இங்கு ஒலியியல் பயன்படுத்தப்படுகின்றன.

எல்லா எழுத்துக்களையும் உச்சரிப்பதை நிறைவேற்ற தமிழ் எழுத்துக்கள் போதுமானவை அல்ல என்று கூற இயலும். மலையாளம், கன்னடா போன்ற தென் திராவிட மொழிகள் உச்சரிப்பு மாற்றங்களைக் கையாள அதிக எழுத்துக்களைக் கொண்டுள்ளன. எனவே தமிழில் உச்சரிப்பு மாற்றங்களைக் கையாள முழு இலக்கண, ஒலி-ஒலியனியல் அமைப்புகளை அறிந்திருப்பது அவசியம். ஆங்கிலத்தில் எழுத்துக்கள் உச்சரிப்பு மயக்கம் காட்டுகின்றன. தமிழில் அத்தகைய நிலை இல்லை.

பின்வரும் வாக்கிய எடுத்துக்காட்டுகளை நோக்குவோம்.

கடவுள் பூமிக்குச் சென்றார்.

கணபதி பள்ளிக்குச் சென்றான்.

முதல் வாக்கியதில் முதல் எழுத்து *க* என்பது [ka] எனவும் இரண்டாவது வாக்கியத்தில் முதல் எழுத்து க எனபது [ga] எனவும் உச்சரிக்கப்படுகின்றது. இவ்விதமான உச்சரிப்பு வேறுபாடுகள் உள்ள எழுத்துக்கள் தமிழில் மேலும் உள்ளன. பின்வரும் எடுத்துக்காட்டுகளை நோக்கவும்.

வீட்டைப் பூட்டிச் சென்றார்.

கடவுள் பூமிக்கு வந்தார்.

முதல் வாக்கியத்தில் பூ என்பது [pU] என்றும் இரண்டாம் வாக்கியத்தில் பூ என்பது [bU] என்றும் உச்சரிக்கப்படுகின்றன. இது போன்று *ச* என்ற எழுத்தும் வேறுபட்டு உச்சரிக்கப்படும். பின்வரும் எடுத்துக்காட்டுகள் இதை தெளிவுபடுத்தும்.

அவன் திசை காட்டினான்.

செடி பச்சையாக இருக்கும்.

முதல் வாக்கியத்தில் சை என்பது [sai] எனவும் இரண்டாவது வாக்கியத்தில் [cai] எனவும் உச்சரிக்கப்படுகின்றது. இந்த உச்சரிப்பு மாற்றத்தை விதிகளால் தர இயலும். ச் என்பது உயிர்களுக்கிடையில் [s] எனவும் *ச* என்பதற்குப் பிறகு [c] எனவும் உச்சரிக்கப்படுகின்றது என விதி தரலாம். *ச* என்பது மொழி முதலில் சில பேச்சு நடைகளில் [sa] எனவும் வேறு சில பேச்சு நடைகளில் [ca] எனவும் உச்சரிக்கப்படும்.

சங்கம் [sanggam] / [canggam]

சாதி [saadi] /[caadi] /[jaadi]

பின்வரும் வாக்கியங்களை எடுத்துக்கொள்வோம்.

ராமன் பாடம் படித்தான்.

அருண் பட்டம் விட்டான்.

முதல் வாக்கியத்தில் ட என்பது [da] என்றும் இரண்டாவது வாக்கியத்தில் [ta] எனவும் உச்சரிக்கப்படுவதை உணரலாம். இந்த உச்சரிப்பு மாற்றத்தை விதிகளால் தர இயலும். என்பது உயிர்களுக்கிடையில் [d] எனவும் ட்-க்குப் பிறகு [t] எனவும் உச்சரிக்கப்படுகின்றது என விதி தரலாம்.

இவை மட்டுமன்றி சில தமிழ் எழுத்துக்கள் ஒத்தறி அடிப்படையில் நீண்ட, குறுகிய, மித உச்சரிப்பு அளவெல்லைகளை அவை வரும் இடத்திற்குத் தகுந்தாற் போல் கொண்டிருக்கும். எடுத்துக்காட்டாக *து* என்பது அது வரும் இடத்திற்குத் தகுந்தாற்போல் ஒத்தறி அடிப்படையில் நீண்ட, குறுகிய, மித உச்சரிப்பைக் கொண்டிருக்கும்.

எ.கா.

பந்து (து நீண்ட உச்சரிப்பைக் கொண்டுள்ளது.)

துவங்கினான் (து குறுகிய உச்சரிப்பைக் கொண்டுள்ளது)

பாதுகாப்பு (து மித உச்சரிப்பைக் கொண்டுள்ளது)

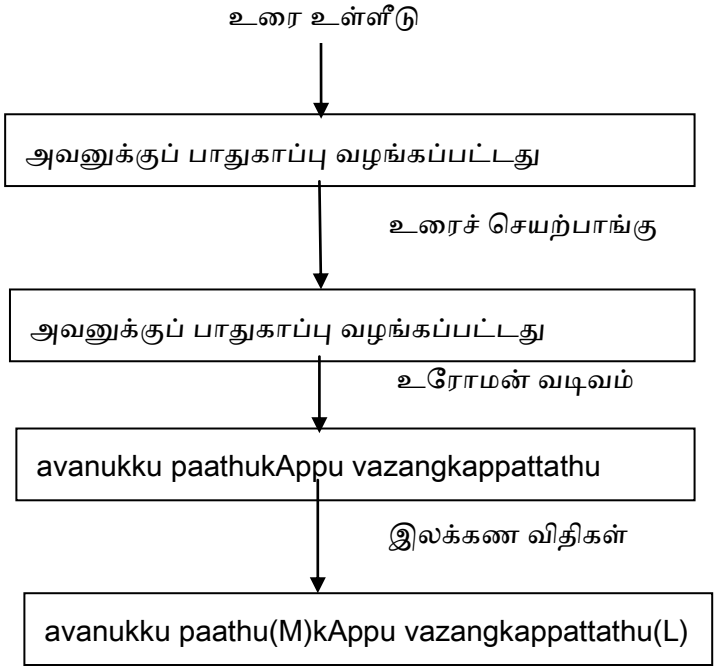
இதுபோன்று பு என்பதும் பின்வருமாறு மூன்று உச்சரிப்புகளைக் கொண்டிருக்கும்.

பருப்பு (பு நீண்ட உச்சரிப்பைக் கொண்டுள்ளது.)

புலவர் (பு குறுகிய உச்சரிப்பைக் கொண்டுள்ளது)

அன்புடையார் (பு மித உச்சரிப்பைக் கொண்டுள்ளது)

60% எழுத்துக்கள் இந்தப் பண்பைக் கொண்டிருக்கின்றன. எனவே பனுவலிலிருந்து பேச்சு கூட்டிணைப்பாக்கத்திற்கு இவ்வகையிலான தகவல்கள் தேவை. நமது வழிமுறைவரைவு எழுத்துக்களின் இடம் அல்லது சூழல் அடிப்படையிலான உச்சரிப்பு மாற்றங்களைக் கையாளும் வகையில் அமைந்துள்ளது. பின்வரும் ஒழுக்குப்படம் பனுவலிலிருந்து பேச்சை கூட்டிணைப்பாக்கம் செய்யும் செயற்பாட்டை ஒரு எடுத்துக்காட்டால் விளக்கும்.



பேச்சுத் தரவு அடித்தளம்

இந்த ஆய்வின் முக்கியமான பகுதி தமிழின் எல்லா எழுத்துக்களையும் அதாவது 247 எழுத்துகளையும் தேவையான ஒரு சில வடமொழி எழுத்துக்களையும் ஒலிப்பதிவு செய்வதாகும்.

இறுதியில் தரவு அடித்தளம் 350 வரை பேச்சுக் கோப்புக்களைக் கொண்டுள்ளது. இலக்கண அடிப்படையில் அதாவது சூழல் அடிப்படையில் உச்சரிப்பில் வேறுபடும் எழுத்துக்கள் அவற்றின் தொடர்புள்ள பேச்சுக் கோப்பில் வைக்கப்படவேண்டும்; எனவே அவைகளும் கோப்புகளில் ஒலிப்பதிவு செய்யப்பட்டிருப்பது அவசியமாகும். பெரும்பாலான தமிழ் எழுத்துக்கள் மூன்று வகையான உச்சரிப்புக்களைக் கொண்டிருக்கும். இத்தகவல் பேச்சு வெளியிட்டுக்கு முக்கியமாகும்.

இயல்பான எழுத்துக்களிலிருந்து ஒரே எழுத்தின் மூன்று வேறுபட்ட வடிவங்கள் ஒலிப்பதிவு செய்யப்பட்டு தரவு அடித்தளத்தில் சேகரிக்கப்படவேண்டும். ஒரே எழுத்து வேறுபட்ட இடங்களில் வேறுபட்டு உச்சரிக்கப்படுவதால் அடிப்படை நெடுங்கணக்கிலிருந்து தனியாக இலக்கண அடிப்படையில் அல்லது சுழல் அடிப்படையில் வேறுபடும் எழுத்துக்கள் தனியாக ஒலிப்பதிவு செய்யப்படவேண்டும். இந்த தரவு அடிப்படையால் நாம் மிக துல்லியமான உச்சரிப்பைப் பெற இயலாது; அதாவது இயல்பான அல்லது இயற்கையான உச்சரிப்பைப் பெற இயலாது.

நாம் எல்லா தமிழ் சொற்களையும் ஒலிப்பதிவு செய்ய இயலும்; இதனால் நாம் மிக உயர்ந்த இயல்பான உச்சரிப்பைப் பெற இயலும்; ஆனால் எல்லாத் தமிழ்ச் சொற்களையும் ஒலிப்பதிவு செய்ய இயலாது. குறைந்தது அடிக்கடி பயன்படுத்தப்படும் சொற்களைத் தரவு அடித்தளத்தில் ஒலிப்பதிவு செய்து வைத்திருப்பது நல்ல இயல்புத்தன்மைக்கு வழி வகுக்கும். தரவு அடித்தளத்தில் இதைச் சேர்ப்பது அதிக அளவு இடம் தேவைபடாது; மட்டுமன்றி; முந்தைய நெறிமுறையோடு ஒப்பிடுகையில் செயற்பாங்கு காலமும் குறைவாகும்.

இருப்பினும் இது முழுச்சொல்லைப் பதிவு செய்வது போன்று வெளியீட்டுப் பேச்சு இயல்புத்தன்மையைத் தரும். அடிக்கடி பயன்படுத்தப்படும் சொற்கள் உருபனியல் அடிப்படையில் திரிபுற்று வரலாம். எனவே இது அதிக காலத்தை எடுக்கும் செயல்பாடாகும். உயிர்கள் மற்றும் மெய்களின் ஒன்றிணைப்பு பனுவல்-பேச்சு ஒழுங்குமுறையில் நல்ல விளைவைத் தரும். இந்த ஒன்றிணைப்பு 4457 கோப்புகளைக் கொண்டிருக்கும்; இது தரவு அடித்தளத்தில் சேர்க்கப்படும். எடுத்துக்காட்டாக ஒன்றிணைப்பு அட்டவணை பின்வருமாறு அமையும்.

உயிரொலிகள் மெய்யொலிகள்	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
க்	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ

ங்	ங											
ச்	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ்	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட்	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண்	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த்	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந்	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப்	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம்	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய்	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர்	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல்	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	வௌ
வ்	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ்	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள்	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற்	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

தரவு அடித்தளப் பகுதி முடிந்துவிட்டால் இலக்கண விதிகளிலிருந்தான வெளியீடு உரோமன் வடிவில் இருக்கும்; இது ஏற்கனவே அசைபிரிக்கும் செயற்பாங்கிற்கு உள்ளாகி இருக்கும். எனவே அலகுத் தேர்வுப் பகுதி சரியான பேச்சுக் கோப்பை ஏற்கனவே உருவாக்கப்பட்ட பேச்சு அடித்தளத்திலிருந்து எடுக்கும். இந்த அலகுத் தேர்வில் செய்யப்படும் முக்கியச் செயற்பாங்கு தொடர்புள்ள அசைபிரிக்கப்பட்ட வாக்கியத்தின் பேச்சுக் கோப்பை எடுக்கும் மற்றும் இந்தப் பேச்சு அலகுகளை குழுமும் அமைப்பை உருவாக்கும்.

ஒன்றிணைப்பாக்கம் (Synthesis)

இறுதியான மற்றும் முடிவான செயற்பாங்கு ஒன்றிணைக்கும் செயற்பாங்காகும். வரிசைப்படுத்தப்பட்ட பேச்சு அலகுகள் ஒன்றிணைப்பு வழிமுறைவரைவைப் பயன்படுத்தி ஒன்றிணைக்கப்படும். பேச்சுக் கோப்புகளின் ஒன்றிணைப்பு மாட்லாபைப் பயன்படுத்திச்

செய்யப்படுகின்றது. ஒன்றிணைக்கும் செயற்பாங்கின் முக்கியமான சிக்கல் சேரும் இடங்களில் தடுமாற்றங்கள் ஏற்படுவதாகும். இவ்வாய்வில் தடுமாற்றங்கள் நேர்த்தியான ஒலிப்பதிவால் தவிர்க்கப்பட்டுள்ளது. ஒன்றிணைப்புச் செயற்பாங்கு அலகுத் தேர்வுச் செயற்பாங்கின் வெளியீடாகத் தரப்படும் எல்லாப் பேச்சுக் கோப்புகளையும் ஒன்றிணைக்கும் மற்றும் ஒரு தனிப் பேச்சுக் கோப்பாக மாற்றும். இதை எப்போது வேண்டுமானாலும் இயக்கவோ நிறுத்தவோ இயலும். இவ்வாய்வின் முக்கியமான நோக்கம் பேச்சு வெளியீட்டில் இயற்கைத்தன்மையை/ இயல்புத்தன்மையைப் பெறுவதாகும்.

முடிவு

உரை-பேச்சு ஒழுங்கமைப்பின் வெளியீட்டைக் காட்டுவதற்கு வரைபட பயன்பாட்டாளர் இடைமுகம் (Graphical User Interface (GUI) தயார் செய்யப்பட்டுள்ளது. ஜி.யு.ஐ.-க்குப் பின்னால் பனுவல் செயற்பாங்கிற்கு ஒரு வழிமுறைவரைவும் ஒன்றிணைப்பாக்கத்திற்கு மற்றொரு வழிமுறைவரைவும் ஆக்கப்பட்டுள்ளன. பனுவல் செயற்பாங்கு வழிமுறைவரைவு பெர்ல் வழியமைப்பால் ஆக்கப்பட்டுள்ளது. ஒன்றிணைப்பு வழிமுறைவரைவு மேட்லாப் (Mat Lab) வழியமைப்பால் ஆக்கப்பட்டுள்ளது. தமிழ் வாக்கியத்தைப் பனுவல் பெட்டியில் தட்டச்சு செய்து பனுவல்-குரல் பொத்தானைச் சுட்டினால் இறுதிப் பேச்சு வெளியீடு கேட்கும். பேச்சுக் கோப்பை நிறுத்தல் பொத்தானைச் சுட்டி எந்தக் காலத்திலும் நிறுத்த இயலும்.

3.6. பேச்சிலிருந்து உரையாக்கம் (speech to text)

பேச்சு அறிதல் (speech recognition) என்பது கணினி மொழியியலின் ஒரு இடைநிலை துணைத் துறையாகும்; இது கணினிகள் மூலம் பேசும் மொழியை உரையாக மாற்றவும் மொழிபெயர்க்கவும் உதவும் முறைகள் மற்றும் தொழில்நுட்பங்களை உருவாக்குகிறது. இது தானியங்கிப் பேச்சு அறிதல் (Automatic Speech Recognition-ASR), கணினிப் பேச்சு அறிதல் அல்லது பேச்சிலிருந்து உரை (Speech to text - STT) என்றும் அழைக்கப்படுகிறது. இது மொழியியல், கணினி அறிவியல் மற்றும் மின் பொறியியல் துறைகளில் அறிவு மற்றும் ஆராய்ச்சியை ஒருங்கிணைக்கிறது.

சில பேச்சு அறிதல் ஒழுங்குமுறைகளுக்கு "பயிற்சி" ("பதிவு" என்றும் அழைக்கப்படுகிறது) தேவைப்படுகிறது, அங்கு ஒரு தனிப்பட்ட பேச்சாளர் உரை அல்லது தனிமைப்படுத்தப்பட்ட சொற்றொகையைக் கணினியில் படிக்கிறார். கணினி நபரின் குறிப்பிட்ட குரலைப் பகுப்பாய்வு செய்து, அந்த நபரின் பேச்சு அறிதலில் நன்றாகப் பயன்படுத்துவதற்குப் பயன்படுத்துகிறது, இதன் விளைவாகத் துல்லியம் அதிகரிக்கும். பயிற்சியைப் பயன்படுத்தாத அமைப்புகள் "பேச்சுபவர்

சுதந்திரமான" (speaker independent) ஒழுங்குமுறைகள் என்று அழைக்கப்படுகின்றன. பயிற்சியைப் பயன்படுத்தும் ஒழுங்குமுறைகள் "பேசுபவர் சார்பு" (speaker dependent) ஒழுங்குமுறைகள் என்று அழைக்கப்படுகின்றன.

குரல் அறிதல் (voice identification) அல்லது பேச்சாளர் அடையாளம் (speaker indentification) காணல் என்பது அவர்கள் சொல்வதைக் காட்டிலும் பேச்சாளரை அடையாளம் காண்பதைக் குறிக்கிறது. பேச்சாளரை அறிதல் ஒரு குறிப்பிட்ட நபரின் குரலில் பயிற்சியளிக்கப்பட்ட அமைப்புகளில் பேச்சை மொழிபெயர்க்கும் பணியை எளிதாக்குகிறது அல்லது பாதுகாப்புச் செயல்பாட்டின் ஒரு பகுதியாக பேச்சாளரின் அடையாளத்தை அறிய அல்லது சரிபார்க்க இது பயன்படுத்தப்படலாம்.

தொழில்நுட்ப கண்ணோட்டத்தில், பேச்சு அறிதல் பல பெரிய கண்டுபிடிப்புகளுடன் நீண்ட வரலாற்றைக் கொண்டுள்ளது. மிகச் சமீபத்தில், ஆழ்ந்த கற்றல் மற்றும் பெரிய தரவுகளின் முன்னேற்றங்களால் இந்தத் துறை பயனடைந்துள்ளது. இத்துறையின் முன்னேற்றம் இந்த துறையில் வெளியிடப்பட்ட ஆய்வுக்கட்டுரைகள் மட்டுமல்லாமல், மிக முக்கியமாக உலகளாவிய தொழில்துறை பேச்சு அறிதல் முறைகளை வடிவமைப்பதிலும் பயன்படுத்துவதிலும் பல்வேறு ஆழமான கற்றல் முறைகளைப் பின்பற்றுவதன் மூலமும் அறியப்படுகின்றது

3.6.1. பேச்சு ஒலிகளும் எழுத்துப்பெயர்ப்பும்

பேச்சிலிருந்து உரையாக்கம் செய்வது ஒலிவடிவில் இருக்கும் பேச்சை எழுத்து வடிவில் மாற்றுவதாகும். ஒலிகள் முதலில் ஒலியன்களாக மாற்றப்படவேண்டும். ஒலியன்கள் எழுத்துக்களாக மாற்றப்படவேண்டும். பேச்சு என்பது தொடர்ச்சியான ஒலிகளின் கோவையாகும். இவ்வொலிக் கோவை ஒலியன்களின் கோவையாக மாற்றப்படவேண்டும். ஒலியன்களின் கோவை உருபன்களின் அல்லது சொற்களின் கோவையாக உருப்படுத்தம் செய்யப்படவேண்டும். சொற்களின் கோவை, சொற்றொடர்களின் கோவையாகவோ வாக்கியங்களின் கோவையாகவோ உருப்படுத்தம் செய்யப்படவேண்டும்.

ஒலியன் என்பது ஒரு கண/குழு ஒலிகளின் குழுமம் ஆகும். ஒரு கண/குழு ஒலிகளை ஒலியனாக உருப்படுத்தம் செய்வது சவாலான செயல்பாடாகும். மொழியியலாளர்கள் அறியாத மொழியை ஆவணப்படுத்த முயலுகையில் அவர்கள் இலக்கு மொழி பேசுபரின் பேச்சைப் புரிந்துகொண்டு தமக்கு இருக்கும் பயிற்சியைப் பயன்படுத்தி அதை எழுத்து வடிவில் எழுதுவர். இது எழுத்துப்பெயர்ப்பு எனப்படும் (transcription). இந்த எழுத்துப்பெயர்ப்பு செய்யப்பட்ட ஆவணத்தைப் பயன்படுத்தி அவர்கள் இலக்கு மொழியின் ஒலியன்களைக் கண்டுபிடிப்பர். இந்த

ஒலியன் கோவைகள் முன்னர் சொன்னதுபோல் சொற்கோவைகளாக உருப்படுத்தம் செய்யப்படவேண்டும். இச்செயல்பாடும் பேச்சிலிருந்து உரையாக்குவதை ஒக்கும்

சொற்களின் உச்சரிப்பை ஆய்வது ஒலியியலின் ஒரு பகுதியாகும். ஒலியனியல் உலக மொழிகளில் பயன்படுத்தப்படும் பேச்சு ஒலிகளின் ஆய்வாகும். நாம் ஒரு சொல்லின் உச்சரிப்பை ஒலிகள் அல்லது கூறுகள் இவற்றை உருப்படுத்தம் செய்யும் குறியீடுகளின் கோர்வையாகும் என்று மாதிரிப்படுத்தலாம். நாம் ஒலிகளை ஆங்கிலம் போன்ற நெடுங்கணக்கு மொழியிலுள்ள எழுத்துக்களுடன் ஒற்றுமையுள்ள ஒலிக் குறியீடுகளால் உருப்படுத்தம் செய்வோம். எடுத்துக்காட்டாக, [ல்] என்பதால் உருப்படுத்தப்படும் ஒலி *ல்* என்ற எழுத்துடன் பொருந்தும். அதேபோல *ப்* என்பதால் உருப்படுத்தப்படும் ஒலி [ப்] என்ற ஒலியுடன் பொருந்தும். ஒலிகளுக்கு எழுத்துக்களைக் காட்டிலும் கூடுதல் வேறுபாடு இருக்கும். மீக்கூறு (prosody) இசைமை (pitch) மற்றும் மாத்திரை (duration) என்பனவற்றை உள்ளடக்கும்.

பேச்சை எழுத்துப்பெயர்ப்பு செய்ய வேண்டி உலக ஒலியியல் கழகத்தார் ஒலிகளின் நெடுங்கணக்கைத் தந்துள்ளனர். இது உலக ஒலியியல் கழகத்தார் நெடுங்கணக்கு (International Phonetic Association Alphabets) என்று அழைக்கப்படுகின்றது. பின்வரும் அட்டவணைகள் மெய்யொலிகள் மற்றும் உயிர் ஒலிகளின் உலக ஒலியியல் கழகத்தார் நெடுங்கணக்குகள்ளாகும்.

உலக ஒலியியல் கழகத்தார் நெடுங்கணக்கு (International Phonetic Association Alphabets)

		bilabial	labio-dental	dental	alveolar	retroflex	palato-alveolar	palatal	velar	uvular	glottal
plosives	uv.	p			t	ʈ		c	k	q	ʔ
	v.	b			d	ɖ		ɟ	g	ɢ	
fricatives	uv.	Φ	f	θ	s	ʂ	ʃ	ç	x	χ	h
	v.	β	v	ð	z	ʐ	ʒ	ʝ	ɣ	ʁ	ɦ
nasals		m	ɱ		n	ɳ		ɲ	ŋ	ɴ	
Semi-vowels	uv.	ʍ									
	v.	w	ʋ		ɹ	ɻ		j			
Rolled / trilled		ʙ			r					ʀ	
Tapped / flapped					ɾ	ɽ					
laterals					l	ɭ		ʎ	L		

lateral-fricatives	uv.				ɬ						
	v.				ɮ						

மேற்கண்ட அட்டவணையில் உலக ஒலியியல் கழகத்தார் நெடுங்கணக்கு (International Phonetic Association Alphabet) சாத்தியமான இடங்களில் தமிழ்மொழிக் குறியீடுகளும் தரப்பட்டுள்ளன. கு=குரலொலி, குஇ=குரலிலா ஒலி.

மெய்யொலிகள் (Consonants)

	ஈரிதழ் குஇ கு	பல்லிதழ் குஇ கு	பல் குஇ கு	நுனியண்ணம்/அண்பல் குஇ கு	நுனியண்ணம் பின்புறம் குஇ கு	நாவளை குஇ கு	இடையண்ணம் குஇ கு	கடையண்ணம் குஇ கு	முன்தொண்டை குஇ கு	குரல்வளை குஇ கு
மூக்கொலி	ம் m		ந் n	ன் n		ண் ɳ	ஞ் ɲ	ங் ŋ		
அடைப்பொலி	ப் p b		த் t d	ற் 		ட் ʈ ɖ	ச் ஜ் c ɟ	க் k g		
உரசொலி		வ் f v		ஸ் s z		ஷ் ʂ				ஹ் h
மருங்கொலி				ல் l		ள் ɭ	ழ் ʎ			
ஆடொலி				ற் r						
வருடொலி				ர் ɾ						

உயிரொலிகள் (Vowels)

	front	central	back
high	i y	ɨ ʉ	ɯ u
	ɪ ʏ		ʊ
middle	e ø	ɜ ə ɵ	ɤ o
	ɛ œ	ɐ ʌ	ɔ
low	æ a		ɑ ɒ

உயிர்நெடில்கள் முக்கால் புள்ளிகளால் அடையாளப் படுத்தப்படும் (எ.கா. i:). மூக்கொலி உயிரொலிகள் உயிர்மேல் டில்டேயால் அடையாளப் படுத்தப்படும் (எ.கா. ã)

	front	central	back
high	இ i		உ u
	ஈ i:		ஊ u:
middle	எ e		ஒ o
	ஏ e:		ஓ o:
low		அ α	
		ஆ α:	

3.6.2. தானியங்கு பேச்சுப் புரிதல் (automatic speech recognition (ASR)

தானியங்கு பேச்சுப் புரிதல் (automatic speech recognition (ASR) ஆய்வு ஒலியியக்கவியல் குறிப்புடன் ஒரு கோர்வைச் சொல்லைப் பொருத்தும் ஒழுங்கமைப்புகளை உருவாக்கி கணினி அடிப்படையில் சிக்கலை அணுகுவதாகும். தானியங்கு பேச்சு அறிதல் (automatic speech understanding (ASU) இந்த நோக்கைச் சொல் அளவில் அல்லாமல் வாக்கியத்தின் ஒருவகையிலான அறிதலை உருவாக்குவதாகும்.

எந்தச் சூழலிலும் எந்தப் பேசுபவராலும் பேச்சின் தானியங்கு ஒலிபெயர்ப்பின் பொதுவான சிக்கல் இதுவரை தீர்வுக்கு அப்பாற்பட்டிருக்கின்றது. அண்மைக் காலங்களில் தானியங்குப் பேச்சுப் புரிதல் தொழில்நுட்பம் சில எல்லைக்குட்பட்ட களங்களில் தெரியும்படி முதிர்ந்துள்ளது. மனித-கணினி ஊடாட்டம் (human-computer interaction) ஒரு முக்கிய பயன்பாடாகும். பல செயல்பாடுகள் காட்சி அல்லது சுட்டு இடைமுகத்தால் (visual or pointing interface) தீர்க்கப்படுகையில் மொழிக் கருத்துப் பரிமாற்றம் பயனுள்ளதாக அமையும் செயல்பாடுகளில் அல்லது தட்டச்சுப் பலகை பொருத்தமாக இராத செயல்பாடுகளில் தட்டச்சுப் பலகையைவிட பேச்சு சிறந்த இடைமுகமாக இருப்பதற்கு திறன் கொண்டுள்ளது. இது பயன்படுத்துபவர் கையாளுவதற்குப் பொருளோ கட்டுப்படுத்துவதற்குக் கருவியோ வைத்திருக்கும் கைகள் அல்லது கண்கள் ஓய்வில்லாதிருக்கும் பயன்பாடுகளையும் உட்படுத்தும். மற்றொரு பயன்பாடுக் களம் தொலைபேசியாகும்; இங்குப் பேச்சுப் பயன்பாடு அழைப்புகளை ஏற்கும் "ஆமாம்" என்பதை அறியும் இலக்கங்களைப் பதிப்பதற்கும் அல்லது அழைப்பை வழிபடுத்தவும் உதவும் சில பயன்பாடுகளில் பேச்சிலாத வரைபட பயனாளி இடைமுகத்தைவிட (graphical user interface) பேச்சும் சுட்டும் இணைந்த பல்மாதிரி இடைமுகம் (multi model interace) திறமையானதாக இருக்கும். இறுதியாகத் தானியக்கப் பேச்சுப் புரிதல் கூறுவதெழுதலுக்குப் பயன்படுத்தப்படுகின்றது; அதாவது ஒரு தனிப் பேசுபரின் நீட்சியுற்ற தன்னுரையின் ஒலிபெயர்ப்பு

பேச்சுத் தொழில் நுட்பத்தின் வேறுபட்ட பயன்பாடுகள் சிக்கல்களில் வேற்பட்ட கட்டுப்பாடுகளை முன்வைக்கின்றது மற்றும் வேறுபட்ட வழிமுறைகளுக்குக் கொண்டு செல்கின்றது. இங்கு நாம் நீண்ட சொற்றொகை தொடர்ச்சியான பேச்சுப் புரிதல் (Large-vocabulary Continuous Speech Recognition (LVRSR) என்பதில் கவனம் செலுத்துவோம். "நீண்ட சொற்றொகை" என்பது ஒழுங்குமுறைகள் கிட்டத்தட்ட 5,000-இலிருந்து 60,000 வரை சொற்களை கொண்டிருக்கும் எனப் பொருள்படும். "தொடர்ச்சியான" என்பது சொற்கள் இயற்கையாக ஒன்றிணைந்து செல்லும் என்பதாகும். இது சொற்கள் இடைவெளியால் பின் தொடரப்படும் தன்மைப்படுத்தப்பட்ட சொல் பேச்சு புரிதல் (isolated-word speech recognition) என்பதுடன் முரண்படும்.

பேச்சுப் புரிதல் என்பது பேச்சுக் குறிகளைச் சொற்களின் வரிசைகளாக ஒலிபெயர்ப்பு செய்யும் சிக்கலுடன் தொடர்புடையது. இன்றைய சூழலில் நன்றாகச் செயலாற்றும் ஒழுங்குமுறைகள் பேச்சின் புள்ளியியல் மாதிரிகளைப் பயன்படுத்துகின்றது. இதன் அடிப்படையில் பேச்சு, மொழி மாதிரியால் உருவாக்கப்படுவதாகக் கருதப்படுகின்றது; இது எல்லா w சொல் கோர்வைகளுக்கும் உற்றுநோக்கப்பட்ட குறிகளிலிருந்து தன்னிச்சையாக Pr(w) என்ற மதிப்பீடுகளையும் f(x|w) என்ற நிகழ்வுத்தகவு அடர்த்திச் செயல்பாட்டால் உருப்படுத்தம் செய்யப்படும் x என்ற குறியில் w செய்தியைக் குறியமாக்கம் (encoding) செய்யும் ஒலியியக்க மாதிரியையும் தருகின்றது. உற்றுநோக்கப்பட்ட ஒலியியக்கக் குறி (observed acoustic signal) தரப்படுகையில் மிகவும் சாத்தியமான சொல் கோர்வையைக் கண்டுபிடிப்பது பேச்சுப் புரிதலின் நோக்கம் ஆகும். பேச்சுக் குறியவிளக்கம்/குறியத்திறவு (decoding) சிக்கலானது x என்ற பேச்சுக் குறி தடப்படுகையில் w என்ற சாத்தியத்தை அதிகரிப்பதையோ அதற்கு ஒப்பாக Pr(w) f(x|w) பெருக்கல் பலனை அதிகரிப்பதை உள்ளடக்குகின்றது.

இந்த ஒழுங்குமுறைகள் அடிப்படையாகக் கொண்டிருக்கின்ற கொள்கைகள் பல ஆண்டுகளாக அறியப்பட்டதாகும்; இது பேச்சுப் புரிதலுக்குத் தகவல் கோட்பாட்டின் (information theory) பயன்பாட்டையும் (application) பேச்சு அடையாளக்குறியின் (speech signal) நிறமாலை உருப்படுத்தத்தின் (spectral representation) பயனையும் (use) குறியத்திறவுக்குச் சக்திவாய்ந்த நிரலாக்கத்தின் (dynamic programming) பயனையும் சூழல் கட்டுண்ட ஒலியியக்க மாதிரிகளின் (context dependent acoustic models) பயனையும் உட்படுத்தும். இந்த உத்திகள்/நுட்பங்கள் பத்து ஆண்டுகளுக்கு முன்பே முன்மொழியப்பட்டாலும் அண்மைக்காலத்தில் விரிவான பேச்சு மற்றும் உரைத் தரவுத்தொகுதி கிடைப்பதன் காரணமாகவும் செயல்படுத்துவதற்குச் சிக்கலான மாதிரிகளையும் வழிமுறைகளையும் அனுமதிக்கும் மேம்படுத்தப்பட்ட செய்முறை காரணமாகவும் கூறத்தக்கவிதத்தில் முன்னேற்றம் ஏற்பட்டுள்ளது. பத்தாண்டுகளுக்கும் முன்னர் இருந்த

தொழிநுட்பத்தின் முன்னேற்ற நிலையுடன் ஒப்பிடுகையில் ஒலியியக்க மாதிரியாகத்தின் முன்னேற்றங்கள் வேறுபட்ட தரவு வகைகள் மற்றும் ஒலியியக்க நிலைமைகள் (கட்டுப்பாடுகள்) பற்றி கிடைக்கப்பெறவேண்டிய நியாயமான செயலாக்கத்தைச் சாத்தியப்படுத்தியுள்ளது.

3.6.3. பேச்சுத் தொழில் நுட்பத்தின் மைல் கற்கள்

பேச்சு அறிதல் ஒழுங்குமுறை பெல் லாபில் 1950களில் முதலில் உருவாக்கப்பட்டது. இந்த ஒழுங்குமுறையில் வேலை ஒரு பேசுபவரிலிருந்து இடைநிறுத்தங்களால் பிரிக்கப்பட்ட இலக்கங்களை (டிஜிட்டல்களை) அறிவதாகும். இந்த ஒழுங்குமுறை அனலாக் எலக்ட்ரானிக்சை பயன்படுத்தி உருவாக்கப்பட்டது; இது கூறப்பட்ட இலக்கத்தின் ஃபார்மெண்ட்கள்/ ஒலிச்செறிவுகள் என்று அழைக்கப்படுகின்ற ஒத்ததிர்வு நிகழ்வெண் உச்சிகளைக் கண்டுபிடித்து (resonant frequency peaks) அறிதலைச் செய்கின்றது. இதன் முரட்டுத்தன்மைக்கு விலக்காக இந்த ஒழுங்குமுறை 98% துல்லியத்தை எய்த முடிந்தது. இது எந்திரம் மனிதப் பேச்சைப் புரியமுடியும் என்ற கருத்துருவை நிரூபித்தது.

ஆய்வு 1960களிலும் 1970களிலும் தொடர்ந்து நடந்தது; இது குறிகைச் செயல்முறையையும் (signal processing) மற்றும் பேச்சு அறிதலை (speech recognition) உருவாக்குவதையும் (அமைப்பொழுங்கு அறிதல் (patten recognition) கவனக்குறிப்பு செய்யும் டிஜிட்டல் கணித்தல் தொழிநுட்பத்தால் ஊக்கப்படுத்தப்பட்டது. பேச்சு ஆய்வின் முக்கியமான பங்களிப்புகள் விரைவான ஃபூயர் மாற்றம் (fast fourier transform (FFT)), செப்ட்ரல் ஆய்வு, நேரியல் முன்னறிவிக்கும் குறியமாக்கம் (Linear Predictive Coding (LPC) இவற்றை உட்படுத்தும்; இது முன்னர் பயன்படுத்தப்பட வடிகட்டி வங்கிகளை (filter banks) இடம்பெயர்த்தது. செயற்கை நரம்பு வலைப்பின்னல்கள் (Artificial Neural Net-works) , டைனமிக் டைம் வார்பிங் (Dynamic Time Warping) மற்றும் ஹிட்டன் மார்கொவ் மாதிரிகள் (Hidden Markov Models) என்பன வெற்றிகரமாகப் பேச்சாய்வில் பயன்படுத்தப்பட்டன. இந்த வழிமுறைகள் பல்வேறு பயன்பாடுகளின் பல செயல்பாடு மற்றும் பயனுள்ள ஒழுங்குமுறையாக விளைந்தது.

1980களிலும் 1990களிலும் ஆய்வுகள் தனியக்கப் பேசு ஒழுங்குமுறைகளைப் (Automatic Speech Recognition Systems (ASR) பேசுபவர் சுதந்திரமான தரவுக் கணங்கள்/குழுமங்கள் (speaker independent data sets), பெரிய சொற்றொகுதிகள் (larger vocabularies) மற்றும் இரைச்சல் வலுவான அறிதல் (noise robust recognition) இவற்றை உட்படுத்திய அதிக கலவைத்தன்மையான/கடினமான செயல்களுக்கு விரிவுபடுத்தக் கவனக்குவிப்பு செய்தது. மெல் நிகழ்வெண் செப்ஸ்ட்ரல் கோஎஃபிஷியண்ட்கள் (Mel frequency cepstal coefficients (MFCC))

போன்ற பண்புக்கூறு பிரித்தெடுப்புக்குப் பேச்சுக்கு ஏற்ப உருவாக்கப்பட்ட குறிகை செயலாக்க நெறிமுறைகளை (speech-tailored signal processing methods) உட்படுத்தி முன்னேற்றம் செய்யப்பட்டது. எச்.எம்.எம். நெறிமுறைகள் (HMM methods) என்பன மேலும் சீராக்கப்பட்டுத் தனியக்கப் பேச்சு அறிதல் நிலைநிற்கும் தொழிநுட்பமாக மாறியது. மேலும் ஆய்வுச் சமூகம் ஆய்வாளர்களுக்கிடையே ஒப்பீடு செய்ய வேண்டி தரவு வசதிசெய்வதற்கும் மென்பொருள் பங்கிடவும் கூடுதல் ஏற்புடையதாக்கப்பட்டது.

TIMIT தரவுத்தொகுதி மற்றும் மூலவள மேலாண்மை தரவுத்தொகுதி (Resource Management Corpus) போன்ற தரமான பேச்சுத் தரவுகள் தொகுக்கப்பட்டு வினியோகிக்கப்பட்டன. மேலும் திறந்த மூலக் குறியங்களைக் கொண்ட பேச்சு மென்பொருள் கருவித்தொகுதிகள் கிடைக்கத்தக்கனவாகச் செய்யப்பட்டது; ஹிட்டன் மார்கொவ் மாதிரிக் கருவிப்பெட்டி (Hidden Markov Toolkit (HTK) என்பது ஒரு எடுத்துக்காட்டாகும். இந்தக் காலம் பல வணிகத் தயாரிப்புகளின் வெளியீடுகளால் முக்கியத்துவம் பெற்றது. ஐபிஎம் வழி குரல் (IBM Via Voice) மற்றும் இயற்கையாகப் பேசுகின்ற டிராகன் ஒழுங்குமுறைகள் (Dragon Systems Naturally Speaking) உள்ளடக்கிய சொல்வதெழுதுதல் கணினி ஒழுங்குமுறை (personal computer dictation systems) என்பன மிகப் புகழ்வாந்தவை.

2000களின் முற்பகுதியில், மீள்நிகழும் செயற்கை நரம்பியல் வலைப்பின்னல்களுடன் (recurrent neural networks) இணைந்த மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகள் போன்ற பாரம்பரிய அணுகுமுறைகளால் பேச்சு அறிதல் ஆதிக்கம் செலுத்தப்பட்டது. இருப்பினும், இன்று, பேச்சு அறிதலின் பல அம்சங்கள் 1997 இல் செப் ஹோக்ரீட்டர் & ஜூர்கன் ஷ்மிதூபரால் (Sepp Hochreiter & Jürgen Schmidhuber, 1997) வெளியிடப்பட்ட தொடர்ச்சியான நரம்பியல் வலையமைப்பான நீண்ட குறுகிய-கால எல்லை நினைவகம் [லாங் ஷார்ட்-டெர்ம் மெமரி (எல்எஸ்டிஎம்) Long short-term memory (LSTM)] என்று அழைக்கப்படும் ஆழமான கற்றல் (deep learning) முறையால் கையகப்படுத்தப்பட்டுள்ளன. எல்.எஸ்.டி.எம் ஆர்.என்.என்.-கள் (LSTM RNNs) மறைந்துபோகும் சாய்வு அளபுச் சிக்கலைத் (vanishing gradient problem) தவிர்க்கின்றன; மேலும் இது இது பேச்சுக்கு முக்கியமான ஆயிரக்கணக்கான தனித்துவமான நேர நடபடிகளுக்கு முன்பு நடந்த நிகழ்வுகளின் நினைவுகள் தேவைப்படுகிற "மிக ஆழமான கற்றல்" செயல்பாடுகளைக் கற்றுக் கொள்ள இயலும். 2007ஆம் ஆண்டில், இணைப்பாளர் காலசார் வகைப்பாட்டால் (Connectionist Temporal Classification (CTC) (சி.டி.சி)) பயிற்சியளிக்கப்பட்ட எல்.எஸ்.டி.எம் சில பயன்பாடுகளில் பாரம்பரிய பேச்சு அறிதலை விஞ்சத் தொடங்கியது.

2015ஆம் ஆண்டில், கூகிளின் பேச்சு அறிதல் சி.டி.சி-பயிற்சி பெற்ற எல்.எஸ்.டி.எம் மூலம் 49% வியத்தகு செயல்திறன் அதிகரித்ததாக கூறப்படுகிறது, இது இப்போது கூகிள் குரல் (Google voice) மூலம் அனைத்து ஸ்மார்ட்போன் பயனர்களுக்கும் கிடைக்கிறது (Wikipedia on Speech Recognition).

2010 ஆம் ஆண்டின் முற்பகுதியில், குரல் அறிதல் (voice recognition) என்றும் அழைக்கப்பட்ட பேச்சு அறிதல் (speech recognition), பேச்சாளர் அறிதலிலிருந்து (speaker recognition) தெளிவாக வேறுபடுத்தப்பட்டது, மேலும் பேச்சாளர் சுதந்திரம் ஒரு பெரிய திருப்புமுனையாகக் கருதப்பட்டது. அதுவரை, அமைப்புகளுக்கு "பயிற்சி" காலம் தேவைப்பட்டது. 1987ஆம் ஆண்டு ஒரு பொம்மைக்கான விளம்பரம் "இறுதியாக, உங்களைப் புரிந்துகொள்ளும் பொம்மை" என்ற கோஷத்தைக் கொண்டிருந்தது (Melanie Pinola. 2011); இருப்பினும் இது ”எந்தக் குழந்தைகள் தங்கள் குரலுக்கு பதிலளிக்க பயிற்சி அளிக்க முடியும்" என்று விவரிக்கப்பட்டது (Wikipedia on Speech Recognition).

2017ஆம் ஆண்டில், மைக்ரோசாப்ட் ஆராய்ச்சியாளர்கள் பரவலாக நன்மதிப்பிடப்பட்ட சுவிட்ச்போர்டு செயல்பாட்டில் உரையாடல் தொலைபேசி உரையை எழுத்துப்பெயர்க்கும் வரலாற்று மனித சமநிலை மைல்கல்லை எட்டினர். பேச்சு அறிதல் துல்லியத்தை மேம்படுத்த பன்முனைப்பான ஆழமான கற்றல் மாதிரிகள் (multiple deep learning models) பயன்படுத்தப்பட்டன. பேச்சு அறிதல் சொல் பிழை வீதம் ஒரே அளவுகோலில் ஒன்றாக செயல்பட்ட 4 தொழில்முறை மனித எழுத்துப்பெயர்பாளர்களைக் (professional human transcribers) காட்டிலும் குறைவாக இருந்தது (Wikipedia on Speech Recognition).

3.6.4. பேச்சுச் செயலாக்கத்தின் வகைகள்

பேச்சு செயலாக்கம் பின்வரும் ஆறு வகைகளாகப் பிரிக்கப்பட்டுள்ளன:

1. பேச்சுக் குறிகையின் மொழிசார் பொருளடக்கத்தின் பகுப்பாய்வைக் கையாளும் பேச்சு அறிதல்
2. பேசுபவரை அடையாளம் காணுவதைக் குறிக்கோளாகக் கொண்ட பேச்சாளர் அறிதல்
3. பேச்சு குறிகைகளின் விரிவாக்கம், எ.கா. இரைச்சல் குறைப்பு
4. பேச்சின் சுருக்கமும் பரப்புதலும்
5. மருத்துவ நோக்கத்திற்காகக் குரல் பகுப்பாய்வு
6. பேச்சு இணைப்பாக்கம்: பேச்சின் செயற்கை உருவாக்கம்

பேச்சுப் புரிதல் என்பது ஒலிவாங்கியால் அல்லது தொலைபேசியால் ஏற்கப்படும் ஒலியியக்கக் குறிகையைச் சொற்களின் கணங்களாக மாற்றும் செயற்பாங்கு ஆகும்.

3.6.5. பேச்சு அறிதலின் வகைகள்

பேசுபவரின் பேச்சு அறிதல் அடிப்படையில் பேச்சைப் பேசுபவர் சார்ந்தது (speaker dependent) மற்றும் பேசுபவர் சாராதது (speaker independent) என்று வகைப்படுத்தலாம்.

பேசுபவர் சார்ந்தது

இவ்வகையிலான மாதிரிகள் ஒரேஒரு நபரின் பேச்சு அமைப்பொழுங்கை அறியும்/புரிந்துகொள்ளும். பேச்சுபவர்சார் மென்பொருள் வழக்கமாகச் சொல்வது எழுதும் மென்பொருளில் (dictation software) பயன்படுத்தப்படுகின்றது; பேசுபவர் சாரா மென்பொருள் வழக்கமாக தொலைபேசிப் பயன்பாடுகளில் பயன்படுத்தப்படுகின்றது. பேச்சுபவர்சார் மென்பொருள் ஒரு தனி நபரின் பேச்சின் தனித்தன்மையான பண்புகளைக் கற்றுக்கொண்டுச் செயலாற்றுகின்றது. புதிய பயன்பாட்டாளர்கள் மென்பொருளுடன் பேசி அதைப் பயிற்றுவிக்க வேண்டும்; இதன்படி கணிப்பொறி அந்த நபர் எவ்வாறு பேசுகின்றார் என்று கற்றுக்கொள்ளும். இதன் பொருள் பேச்சுப் புரிந்துகொள்ளும் மென்பொருளைப் பயன்படுத்துவதற்கு முன் பயனர்கள் கணிப்பொறிக்கு அதாவது மென்பொருளுக்குச் சில பக்கங்களைப் படித்துத் தமது பேச்சைக் கணினிக்கு அறிமுகப்படுத்தவேண்டும்.

பேசுபவர் சாராதது

பேசுபவர் சாரா மாதிரிகள் ஒரு பெரிய குழுவைச் சார்ந்த பேசுபவர்களின் பேச்சு அமைப்பொழுங்கை புரிந்துகொள்கின்றன/அறிந்துகொள்கின்றன. இவை எந்த நபரின் குரலையும் புரிந்துகொள்ளும்படி வடிவமைக்கப்பட்டுள்ளன; எனவே பயிற்சி தேவையில்லை. எதிர்மறையான தகவல் என்னவென்றால் பேசுபர் சாரா மென்பொருள் பெரும்பாலும் பேசுபவர் சார்ந்த மென்பொருளைவிட குறைந்த துல்லியம் உள்ளதாய் இருக்கும்.

பேச்சு அறியும் ஒழுங்குமுறைகளை அறியவேண்டிய கூற்றுக்களின் வகையின் விளக்கத்தால் பல வகுப்புகளாகப் பிரிக்கலாம். இந்த வகுப்புகள் எப்போது பேசுபவர் ஒரு கூற்றைத் தொடங்கி முடிக்கிறர் என்பதைப் புரிந்துகொள்ளும் பேச்சுப்புரியும் ஒழுங்குமுறையின் திறன் அடிப்படையில் அமையும். பின்வருவன சில வகைகளாகும்.

தனிச் சொற்கள் புரிவான்கள்

தனிச்சொற்கள் புரிவான்கள் (isolated word recognizers) வழக்கமாகப் பதமாதிரியின் (sample) இருபக்கங்களிலும் மௌனத்தை (கேட்புக் குறிகைகளின் குறைவு) வேண்டும். இதன்

அர்த்தம் இது தனிச் சொற்களை மட்டும் ஏற்கும் என்பதல்ல; இது ஒருநேரத்தில் ஒரு கூற்றை வேண்டும் என்பதாகும். பெரும்பாலும் இந்த ஒழுங்குமுறைகள் 'கேள்/கேட்காதே' நிலைகளைக் கொண்டிருக்கும்; அவை கூற்றுக்களுக்கிடையில் பேசுபவர் காத்திருப்பதை வேண்டும்.

இணைக்கப்பட்ட சொற்களுக்கான பேச்சு ஒழுங்குமுறைகள்

இணைக்கப்பட்ட சொற்களுக்கான பேச்சு ஒழுங்குமுறைகள் (connected word recognition system) தனிச்சொற்கள் ஒழுங்குமுறையைப் போன்றவை; ஆனால் தனிக் கூற்றுக்களை அவற்றிற்கிடையே குறைந்த இடைநிறுத்தத்துடன் சேர்ந்து போக அனுமதிக்கின்றது.

தொடர்ச்சியான பேச்சுக்கான பேச்சு அறிதல் ஒழுங்குமுறைகள்

தொடர்ச்சியான பேச்சுக்கான பேச்சு அறிதல் ஒழுங்குமுறைகள் (Continuous speech recognition system) சிக்கலானது; ஏனென்றால் கூற்று எல்லைகளை நிர்ணயிக்கச் சிறப்பான நெறிமுறைகளைப் பயன்படுத்தவேண்டும். தொடர்ச்சியான பேச்சு அறிவான்கள் (continuous speech recognizers) பேசுபவரை இயற்கையாகப் பேச அனுமதிக்கின்றது; கணினி பொருளடக்கத்தை நிர்ணயிக்கும்.

தன்னியல்பான பேச்சு பேச்சு அறிதல் ஒழுங்குமுறைகள்

தன்னியல்பான பேச்சுக்கான பேச்சு அறிதல் ஒழுங்குமுறைகள் (spontaneous speech recognition systems) சொற்கள் மேலூர்வது, 'உம்', 'ஆஹா' போன்ற உணர்வுக்குறிகள் மற்றும் சிறிய விக்கல்கள் போன்ற பலவிதமான இயற்கையான பேச்சுப் பண்புக்கூறுகளைக் கையாளத் திறன் கொண்டிருக்கவேண்டும். இந்த ஒழுங்குமுறைகள் பேச்சை இயற்கையான/இயல்பான சூழலில் அறியவேண்டும்; பேச்சு ஒத்திகைப் பார்த்துபோன்று செயற்கையாக இருக்காது.

குரல் சரிபார்பு அல்லது அடையாளம் காணுதல்

இவ்வகையிலான பேச்சு அறிவான் குறிப்பிட்ட பேசுபவரை அடையாளம் காணும் திறன் கொண்டிருக்கவேண்டும்.

3.6.6. பேச்சு அறிதல் ஒழுங்குமுறைகளின் பயன்களும் பயன்பாடுகளும்

பேச்சு அறிதல் ஒழுங்குமுறைகளின் பயன்களும் பயன்பாடுகளும் பலவாகும். பின்வருவன மிகப் பொதுவானவைகளாகும்.

சொல்வதெழுதுதல் (dictation)

சொல்வதை எழுதும் பேச்சு அறியும் ஒழுங்குமுறை மிகப் பொதுவான பயன்பாடாகும். இது மருத்துவம்சார் ஒலிபெயர்த்தல் (medical transcription) சட்டம்சார் மற்றும் வணிகச் சொல்வதெழுதுதல் மற்றும் பொதுவான சொல் பகுப்பாய்வு இவற்றை உள்ளடக்கும். சில

நேர்வுகளில் ஒழுங்குமுறையின் துல்லியத்தை அதிகரிக்கச் சிறப்புச் சொற்றொகுதிகள் பயன்படுத்தப்படும்.

கட்டளையும் கட்டுப்பாடும்

செயல்பாடுகளையும் செயல்களையும் நிகழ்த்த வேண்டி வடிவமைக்கப்பட்ட ஒலியியக்கப் பேச்சுப் புரிதல் (Acoustic Speec Recognition (ASR)) ஒழுங்குமுறைகள் கட்டளை மற்றும் கட்டுப்பாடு ஒழுங்குமுறைகள் (Command and control systems) என்று விளக்கம் பெறும். “Open Netscape”, “Start a new xterm” போன்ற கூற்றுகள் இதைச் செய்யும்.

தொலைபேசி

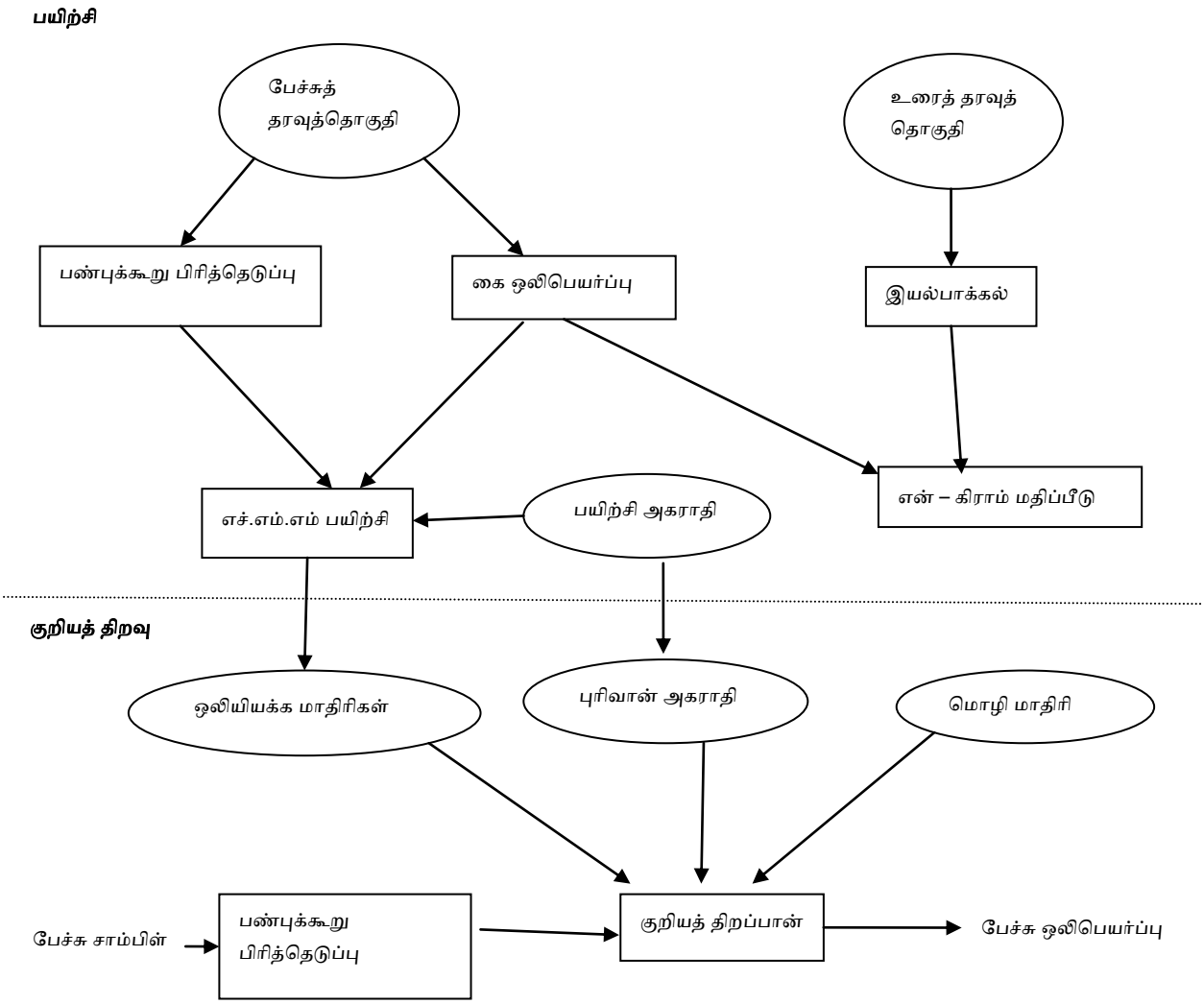
PBX அல்லது குரல் அஞ்சல் (Voice Mail) ஒழுங்குமுறைகள் குறிப்பிட்ட சுரத்தை அனுப்ப வேண்டி பொத்தானை அழுத்துவதற்குப் பதிலாக அழைப்பவர்கள் கட்டளைகளைப் பேச அனுமதிக்கின்றது.

மருத்துவ ஊனங்கள்

திரும்பநேரும் மிகைமுயற்சிக் காயங்கள் (repetitive strain injuries (RSI), தசைத் தேய்வு (muscular dystrophy) மற்றும் பிற உடல்சார்ந்த எல்லைகள் காரணமாகப் பலருக்கும் தட்டச்சுசெய்வது கடினமாக இருக்கும்; இவர்கள் பேச்சு புரிதல் ஒழுங்குமுறையைப் பயன்படுத்துகின்றனர். பேசைக் கேட்பதில் சிரமம் உள்ளவர்கள் தொலைபேசியுடன் இணைக்கப்பட்ட ஒழுங்குமுறையைப் பயன்படுத்தலாம்; அது அழைப்பவர்களின் பேச்சை உரையாக மாற்றித்தரும்.

3.6.7. பொதுவான பேச்சுப்புரிதல் ஒழுங்குமுறை

ஒரு பொதுவான பேச்சுப் புரிதல் ஒழுங்குமுறையின் (generic specch recognition system) முக்கிய உட்கூறுகள் பின்வரும் படத்தில் காட்டப்பட்டுள்ளது (Lamel and Gauvain, 2003: 306). முக்கிய அறிவு மூலங்கள் (பேச்சு மற்றும் உரை பயிற்சிப் பொருட்கள் மற்றும் உச்சரிப்பு அகராதி), பண்புகூறு ஆய்வு (அல்லது அளபுருவாக்கம் (parameterization)), பயிற்சிக் கட்டத்தில் மதிப்பிடப்படும் ஒலியியக்க மற்றும் மொழி மாதிரிகள் மற்றும் குறியத்திறபான்கள்/குறிய விளக்கிகள் (decoders) என்ற தனிமங்கள் காட்டப்பட்டுள்ளன. இவற்றின் விளக்கம் கீழே தரப்பட்டுள்ளது (Lamel and Gauvain, 2003: 306).



3.6.8. ஒலியியக்க அளபுருவாக்கமும் மாதிரியாக்கமும்

ஒலியியக்க அளபுருவாக்கமும் மாதிரியாக்கமும் (acoustic parameterizagtion and modelling) பேச்சுப் புரிதலுக்கான மொழியியல் தகவல்களைப் பராமரிக்க முயலுகையில் மாதிரிச் சிக்கலைக் குறைப்பதற்காக ஒலியியக்கப் பண்புக்கூறுகளின் தேர்வு (choice) மற்றும் உகப்பாக்கத்துடன் (மிகையாக்கம்) (optimization) அக்கறைகொண்டுள்ளது. ஒலியியக்க மாதிரியாக்கம் பேச்சு அடையாளக்குறியில் இருக்கும் மாறுபடும்தன்மையின் வேறுபட்ட மூலங்களைக் கணக்கில் எடுத்துக்கொள்ளவேண்டும்: மொழியியல் சூழலில் தோன்றுபவை மற்றும் பேசுபவர் மற்றும் ஒலியியக்கச் சூழல் (எ.கா. பின்னணி இரைச்சல், இசை) போன்ற மொழியியல் அல்லாத சூழல் மற்றும் பதிவு அலைவரிசை (recording channel) (எ.கா. நேரடியான ஒலிபெருக்கி, தொலைபேசி) இவற்றுடன் தொடர்புடையவை. தற்போதைய தொழில்நுட்ப

ஒழுங்குமுறைகளானது (state-of-art systems) ஒலியியக்கப் பண்புக்கூறு திசையன்களின் (acoustic feature vectors) ஒரு கோர்வையின் நிகழ்தகவு அடர்த்திச் செயற்பாட்டை (probability density function) மாதிரியாக்கம் செய்வதைக் கொண்டிருக்கும் ஒலியியக்க மாதிரியாக்கதிற்கு மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளைப் (Hidden Markov Models (HMM)) பயன்படுத்துகின்றது. இங்கு பொதுவான அளபுருவாக்கங்கள் விளக்கப்படுவதுடன் ஒலியியக்க மாதிரி மதிப்பீடு/கணக்கீடு (acoustic model estimation) மற்றும் ஏற்றுக்கொள்கை (adaptation) விவாதிக்கப்படும் (Lamel and Gauvain, 2003: 307-312).

3.6.8.1. ஒலியியக்கப் பண்புக்கூறு ஆய்வு

ஒலியியக்கப் பண்புக்கூறிகளின் ஆய்வின் முதல் படி இலக்கமாக்கல் (digitization) ஆகும்; இதில் தொடர்ச்சியான பேச்சு அடையாளக்குறி வெவ்வேறான மாதிரிக்கூறுகளாக மாற்றப்படும். மிகப் பொதுவாகப் பயன்படுத்தப்படும் மாதிரிக்கூறு வீதங்கள் நேரடியான ஒலிவாங்கி உள்ளீட்டிற்கு 16 கி.ஹெட்சும் 10 கி.ஹெட்சும் ஆகும்; தொலைபேசி அடையாளக்குறிகளுக்கு 8 கி.ஹெட்ஸ் ஆகும். அடுத்த படி பண்புக்கூறு பிரித்தெடுப்பாகும் (அளபுருவாக்கம்அல்லது முன்னிறுதி ஆய்வு); முக்கியமான மொழியியல் தகவல்களைத் தக்கவைத்துகொண்டு மிகைமையை நீக்கவும் மாறுபடும்தன்மையை குறைக்கவும் முயற்சிசெய்துகொண்டு கூடுதல் கச்சிதமான வழியில் கேட்பொலி அடையாளக்குறிகளை உருப்படுத்தம் செய்வது அதன் நோக்கமாகும். பெரும்பாலான புரிதல் ஒழுங்குமுறைகள் ஃபூயர் மாற்றம் (Fourier transform) அல்லது நேரியல் முன்கணிப்பு மாதிரி (linear prediction model) அடிப்படையிலான குறுகிய கால செப்டரல் பண்புக்கூறுகளைப் (short-time spectral features) பயன்படுத்துகின்றன. கச்சிதமான உருப்படுத்தம் என்பதாலும் நேரடி செப்டரல் கூறுகளைக்காட்டிலும் (direct spectral components) குறைவாகவே இயைபுபடுத்தப்பட்டுள்ளது என்ற காரணத்தாலும் செப்டரல் அளகைகள் பிரபலமானவை ஆகும். இது பண்புக்கூறு சார்பை (feature dependency) மாதிரிப்படுத்த வேண்டிய தேவையை குறைத்து எச்.எம்.எம். அளகைகளின் (HMM parameters) கணக்கீட்டை எளிமைப் படுத்துகின்றது. பேச்சுக் குறிகள் தொடர்ச்சியாக மாறிக்கொண்டிருந்தாலும் பேச்சொலிப்பான்கள் இயங்கவியலும் விகிதத்தின் மீதான இயற்பியல்சார் கட்டுப்பாடுகள் காரணமாகக் குறிகளைக் குறைந்த கால அளவுகளில் பகுதி நிலையானவை (quasi-stationary) என்று கருத இயலும் என்பது இயல்பான ஊகம் ஆகும்.

மிகப் பிரபலமான இரண்டு குழுமப் பண்புக்கூறுகள் மெல் நிகழ்வெண் செப்ட்ரல் (Mel Frequency Ceptral (MFC) பகுப்பாய்வால் அல்லது பெர்-செப்ட்ரல் நேரியல் முன்கணிப்பு (Per-ceptral Linear Prediction (PLP) பகுப்பாய்வால் கிடைக்கப்பெறும் செப்ட்ரம் குணகங்கள் (ceptrum

coefficients) ஆகும். இந்த இரண்டு நேர்வுகளிலும் மெல் அளவு குறைந்தகால திறன் அலைமாலை (short-term power spectrum) ஒரு குறிப்பிட்ட சாளரத்தில் (பெரும்பான்மையும் 20 முதல் 30 மில்லிசெகண்ட்) கணிக்கப்பட்டுள்ளது. ஒலிக்குறியின் சாளரமாகத்தால் ஏற்பட்ட தொடர்பின்மைகள் காரணமாக அலைமாலையில் போலியான மிக உயர்ந்த நிகழ்வெண்ணைத் தவிர்க்க ஹாமிங் சாளரம் (Hamming window) போன்ற ஒடுங்கிய சாளரத்தைப் பயன்படுத்துவது பொதுவானதாகும். சாளரம் பின்னர் இடம்பெயர்க்கப்படும் (பெரும்பான்மையும் சாளர பரிமாணத்தின் முக்கால் அல்லது அரை) மற்றும் அடுத்த பண்புக்கூறு திசையன் (வெக்டர்) கணிக்கப்படும். மிகப் பொதுவாகப் பயன்படுத்தப்படும் ஆப்செட் (நஷ்ட ஈடு/பக்கக்கிளை) 10 மில்லி செகண்ட் ஆகும். குறைந்த நிகழ்வெண் பரப்பெல்லையில் (1000 ஹெட்ஸ்க்கும் கீழ்) மற்றும் 1000க்கும் மேல் லாக்கிரதமிக்காகவும் நேரியலாக இருப்பதால் மெல் அளவு மனித செவிப்புலன் ஒழுங்கமைப்பை (human auditory system) கிட்டத்தட்ட ஒத்திருக்கின்றது. வடிகட்டிவங்கி அளகைகளின் லாக்கின் மறுதலை மாற்றத்தை எடுத்துக்கொண்டு செப்டல் அளகைகள் பெறப்படுகின்றது. எம்.எஃப்.சி. குணகத்தின் நேர்வில் ஒரு கொசைன் மாற்றம் லாக் பவர் அலைமாலைக்குப் பயன்படுத்தப்படுகின்றது; ஆனால் ஒரு ரூட்-லினியார் முன்கணிப்பு குறியனாக்கம் (root-Linear Predictive Coding (LPC) பகுப்பாய்வு பி.எல்.பி. செப்ட்ரம் குணகங்களைப் பெறுவதற்காகப் பயன்படுத்தப்படுகின்றது. LVCAR-க்கு இரு குழுமப் பண்புக்கூறுகளும் வெற்றிகரமாகப் பயன்படுத்தப்படுகின்றது; ஆனால் பி.எல்.பி. பகுப்பாய்வு பின்னணி இரைச்சலின் இருப்பில் சில ஒழுங்குமுறைகள் கூடுதல் பலமானதாகக் காணப்படுகின்றது. ஒலிக் குறியின் சாளரப் பகுதியுடன் தொடர்புடைய செப்டல் குணகத்தின் குழுமம் சட்டகம் அல்லது பாரமீட்டர் வெக்டர் என்று குறிப்பிடப்படுகின்றது. செப்டரல் சராசரியைக் கணிப்பது செயற்பாங்கிற்கு முன்னரே ஒலிக்குறிகள் எல்லாம் கிடைப்பதை வேண்டும்; செயற்பாங்கு பதிவுசெய்வதுடன் ஒத்திசையவேண்டி வருகிற ஒருசில பயன்பாட்டில் இது அப்படியல்ல. இந்நேர்வில் செப்ட்ரல் குறைத்தலில் மாற்றியமைக்கப்பட்ட வடிவதை நிறைவேற்ற இயலும்; இதில் இறுதி n சட்டகங்களிலிருந்து (n என்பது பேச்சின் 1sக்கு இணையாக/பொருத்தமாகக் கிட்டத்தட்ட 100 எனத் தெரிந்தெடுக்கப்படும்) நடப்பு சராசரி கணக்கிடப்படுகின்றது. பேச்சுக் குறிகளின் இயக்க இயல்பைப் பெறுவதற்காக 'டெல்டா' அளகை கொண்ட பண்புக்கூறு திசையனை அதிகரிப்பது பொதுவானதாகும். அடுத்தடுத்த சட்டகங்களில் அளகைகளின் முதல் மற்றும் இரண்டாவது வேறுபாடுகளை எடுத்துக்கொண்டு டெல்டா அளகைகள் கணக்கிடப்படுகின்றது.

3.6.8.2. ஒலியியக்க மாதிரிகள்

மறைக்கப்பட்ட மார்கோவ் மாதிரிகள் ஒலியியக்கப் பண்புக்கூறு திசையன்களின் (acoustic feature vectors) கோர்வைகளை மாதிரிப்படுத்தப் பரவலாகப் பயன்படுத்தப்படுகின்றது. இம்மாதிரிகள் அவை செயல்திறன்மிக்கவை மற்றும் நன்றாக நிறுவப்பட்ட நுட்பங்களைப் பயன்படுத்தி அவற்றின் அளகைகளைத் திறமையாக மதிப்பிட இயலும் என்ற காரணங்களால் பிரபலமானவை. இரண்டு கட்டங்களில் பேச்சுப் பண்புக்கூறு திசையன்களின் (speech feature vector) உற்பத்தியை மாதிரிப்படுத்த அவை பயன்படுத்தப்படுகின்றன. முதலாவது மார்கோவ் சங்கலி ஒரு கோர்வை நிலைகளை உருவாக்கப் பயன்படுத்தப்படுகின்றது; இரண்டாவது ஒவ்வொரு நிலையுடன் தொடர்புடைய நிகழ்தகவு அடர்த்திச் செயல்பாட்டைப் (probability density function) பயன்படுத்திப் பேச்சுத் திசையன்கள் பெறப்படுகின்றன. மார்க்கோவ் சங்கலி நிலைகளின் எண்ணிக்கையாலும் நிலைகளுக்கு இடையிலுள்ள நிலைமாற்ற நிகழ்தகவுகளாலும் விளக்கப்படுகின்றது.

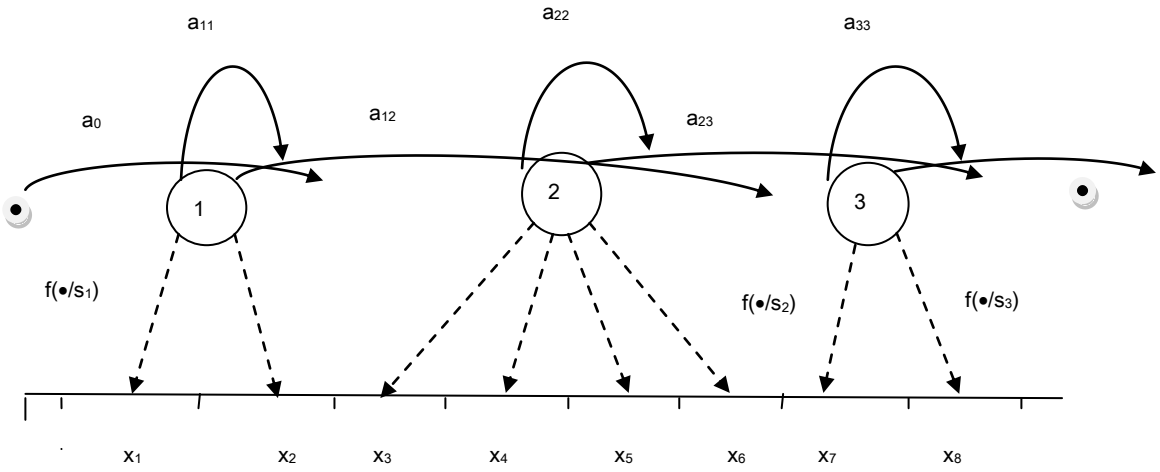
எல்.வி.சி.எஸ்.ஆர் (LVCSR) ஒழுங்குமுறைகளில் மிகப்பரவலாகப் பயன்படுத்தப்படும் தொடக்கநிலை ஒலியியக்க அலகுகள் ஒலி அடிப்படையிலானதாகும்; இதில் ஒவ்வொரு ஒலியும் குறைந்த எண்ணிக்கையிலான நிலைகள்கொண்ட மார்கோவ் சங்கிலியால் உருப்படுத்தம் செய்யப்பட்டுள்ளது. வேறுபட்ட கட்டமைப்பியல்கள் முன்மொழியப்பட்டாலும் காலக்குறுக்காக அலைமாலை மாற்றத்தைப் பெறுவதற்கு இடமிருந்து வலமான நிலைக் கோர்வைகளைப் (state sequences) பயன்படுத்துகின்றது (Lamel and Gauvain, 2003: 309). மிக அதிகமாகப் பயன்படுத்தப்படுகின்ற கட்டமைப்பு ஒவ்வொரு மாதிரிக்கும் உமிழும் நிலைகளை மூன்றிற்கும் ஐந்திற்கும் இடையில் கொண்டிருப்பதாகும்; இங்கு நிலைகளின் எண்ணிக்கை அலகிற்கு ஒரு குறைந்த கால அளவை வற்புறுத்துகிறது. சில கட்டமைப்புகள் தேவைப்படுகின்ற கால அளவை குறைப்பதற்காகச் சில நிலைகளைத் தவிர்க்க அனுமதிக்கின்றது. கவனித்தலின் நிகழ்தகவு (அதாவது பேச்சு திசையன்) முதல் ஒழுங்குமுறை மார்க்கோவ் கருதுகோள் (firs-order Markov assumption) என்று அழைக்கப்படுகின்ற நிலையைச் சார்ந்திருப்பதாகக் கருதப்படுகின்றது.

சரியாகச் சொன்னால் λ அளகை திசையன் கொண்ட n நிலை எச்.எம்.எம். தரப்படுகையில், எச்.எம்.எம். வாய்ப்பியல் செயல்பாடு பின்வரும் இணைந்த நிகழ்தகவு அடர்த்தி செயல்பாடு $f(x,s|\lambda)$ $x=(x_1,\ldots,x_T)$ என்ற கவனிக்கப்பட்ட ஒலிக்குறி $s = (s_0,\ldots s_T)$ என்ற கவனிக்கப்படாத கோர்வை,

$$f(x,s|\lambda)\ x=\pi_{so}\prod_{t=11}^{T} a_{s\text{-}s}\ f\ x_t|s_t)$$

இதில் π_i தொடக்க சாத்திய நிலை i, a_{ij} i நிலையிலிருந்து j நிலைக்கு நிலைமாற்ற சாத்தியம் மற்றும் $f(.|s)$ என்பது ஒவ்வொரு s நிலையுடன் தொடர்புடைய உமிழும் PDF. பின்வரும் படம் நிலைமாற்ற நிகழ்தகவுகள் மற்றும் பி.டி.எஃப்கள் கவனித்தல் தொடர்புடைய மூன்று எச்.எம்.எம். மூன்று நிலையைக் காட்டுகின்றது (Lamel and Gauvain 2003:309).

பேச்சு கூறுகளை உருப்படுத்தம் செய்யும் (x1…x2) திசையன்களை உருவாக்கும் தவிர்ப்பு நிலை கொண்ட எச்.எம்.எம். ஒரு எடுத்துக்காட்டான இடவல 3-நிலை ஒலி



ஒலி அடிப்படையான மாதிரிகள் (phone based models) ஒரு தரப்பட்ட மொழியின் தொடக்க அலகுகளைப் (elementary units) பயன்படுத்தி அறிதல் அகராதிகளை (recognition dictionaries) விளக்க இயலும் என்ற அனுகூலத்தைத் தருகின்றது; மற்றும் பல மொழியியல் ஆய்வுகளிலிருந்து நன்மையைப் பெறுகின்றது. ஒரு ஒலியனியல் அகராதியைப் பயன்படுத்தாமலேயே "சொல் மாதிரிகள்" ("word models") அல்லது ஃபெனோன்கள் (fenones) போன்ற வேறுபட்ட பொருத்தத்தால் பேச்சு அறிதலைச் செய்ய இயலும். (சொற்கள், அசைகள், உப அசைகள் போன்ற) பெரிய அலகுகளுடன் ஒப்பிடுகையில் சிறிய துணைச்சொல் அலகுகள் அளகைகளின் எண்ணிக்கையைக் குறைக்கின்றது; குறுக்குச்சொல் மதிரியாக்கத்தை இயலச்செய்கின்றது; புதிய சொற்றொகையை எடுத்துச்செல்ல வசதிசெய்கின்றது; மிக முக்கியமாக அரிதான சூழல்களை மாதிரிப்படுத்த பின்வாங்கும் நுட்பங்களுடன் (back-off mechanisms) தொடர்புபடுத்த இயலும். ஃபெனொன் தானியங்கு பயிற்சிக்குக் கூடுதல் அனுகூலத்தைத் (நன்மையை) தர இயலும்;

இருப்பினும் காரணகாரியமான மொழியியல் அறிவை (a priori linguistic knowledge) உட்படுத்த திறனில்லை.

ஒரு தரப்பட்ட எச்.எம்.எம். ஒரு ஒலியின் அண்மையிலுள்ளவைகளை (சூழல்-சாரா அல்லது ஒற்றையொலி மாதிரி (context-independent or monophone model) அல்லது ஒரு குறிப்பிட்ட சூழலில் ஒரு ஒலியைக் (context-dependent model) கருத்தில் கொள்ளாமல் ஒரு ஒலியை உருப்படுத்தம் செய்ய இயலும். சூழல் ஒரு சொல்லுக்குள் (சொல்-இருப்பிடச் சார்பு) ஒரு ஒலியின் இருப்பிடத்தை உட்படுத்தவோ உட்படுத்தாதிருக்கவோ செய்யலாம் மற்றும் சொல் அக மற்றும் குறுக்குச் சொல் சூழல்கள் ஒன்றிணைக்கப்படவோ அல்லது தனியான மாதிரிகளாகக் கருதப்படவோ செய்யலாம். சொல் குறுக்கான சூழல்களைப் பயன்படுத்துவது குறித்திறவைச் சிக்கலாக்கும். நிகழ்வெண் அடிப்படையில் அல்லது கொத்தாக்க நுட்பங்களை (clustering techniques) அல்லது தீர்மானக் கிளை அமைப்புகளைப் (decision trees) பயன்படுத்தி சூழல் அலகுகளைத் தேர்ந்தெடுக்க வேறுபட்ட அணுகு முறைகள் பயன்படுத்தப்படுகின்றன; மற்றும் வேறுபட்ட சூழல் வகைகள் துருவி ஆயப்பட்டுள்ளன: (உள்ளுக்குள் அல்லது குறுக்குச் சொல்) இருப்பிடச் சார்பு கொண்ட மற்றும் சார்பில்லாத தனி ஒலிச் சூழல்கள் (single phone contexts), மூவொலிகள் (triphones), பொதுமையாக்கப்பட்ட மூவொலிகள் (generalilzied triphones), நாலொலிகள் (quadphones), ஐந்தொலிகள் (quinphones). மாதிரி அளவைக் குறைக்கும் படி மாதிரி நிலைகள் கொத்தாக்கம் செய்யப்பட்டன; இது 'கட்டப்பட்ட நிலை' மாதிரிகள் ('tied state models) என்று குறிப்பிடும்படி விளைவை ஏற்படுத்தியுள்ளது.

ஒலியியக்க மாதிரிப் பயிற்சி ஒவ்வொரு எச்.எம்.எம். அளகைகளை மதிப்பீடு செய்வதைக் கொண்டிருக்கும். தொடர்ச்சியான அடர்த்திக்கு காசியன் கலவை (Gaussian mixture) எச்.எம்.எம்.களுக்கு இது சராசிரிகளை/வழிகளை மற்றும் இணைமாறுபாட்டு அணிகள், கலவை எடைகள் மற்றும் நிலைமாற்ற நிகழ்தகவுகள் இவற்றை வேண்டும். மிக பிரபலமான அணுகுமுறைகள் மாதிரிக்கும் பயிற்சித் தரவுக்கும் நல்ல பொருத்தத்தை உறுதிசெய்கின்ற அதிகச் சாத்திய அளவுகோலைப் (maximum likelihood criterion) பயன்படுத்துகின்றன.

மாதிரி அளகைகளின் மதிப்பீடு எதிர்பார்ப்பு உச்சமடையச்செய்தல் (expectation-maximization (EM)) வழிமுறையுடன் செய்யப்படுகின்றது; இது மாதிரி அளகைகளின் தொடக்கக் குழுமத்துடன் தொடங்குகிற பன்முறைநிகழும் செயல்முறையாகும். மாதிரி நிலைகள் பயிற்சித் தரவுக் கோர்வைகளுடன் வரிசைப் படுத்தப்படுகின்றது; மற்றும் இந்த அளகைகள் பாம்-வெல்ச் மீள்மதிப்பீட்டு வாய்ப்படுகளைப் (Baum-Welch re-estimation formulas) (Baum et al 1970; Liporace 1982; Juang 1985) பயன்படுத்தி இப்புதிய வரிசைப்படுத்தல் அடிப்படையில்

மீள்மதிப்பீடு செய்யப்படுகின்றது. இவ்வழிமுறை தரப்பட்ட பயிற்சித் தரவு மாதிரிகளின் நிகழ்வியல்பு ஒவ்வொரு மறுசெய்கையிலும் அதிகரிப்பதை உறுதி செய்கின்றது. வரிசைப் படுத்தும் கட்டத்தில் தரப்பட்ட பேசு சட்டகம் முன்னோக்கிய-பின்னோக்கிய வழிமுறைகளைப் (forward-backward algorithm) பயன்படுத்தி பல்நிலைகளுக்கு அல்லது விட்டர்பி வழிமுறையைப் (Viterbi algorithm) பயன்படுத்தி ஒரு தனி நிலக்கு ஒதுக்கீட்டு செய்ய இயலும். இந்த இரண்டாவது அணுகுமுறை சிறிது குறைந்த நிகழ்வியல்புகளைத் தருகின்றது; ஆனால் நடைமுறையில் குறிப்பாக அதிக அளவிலான தரவு கிடைக்கையில் துல்லியத்தில் அவ்வளவாக வேறுபாடு இருக்காது. எதிர்பார்ப்பு உச்சமடையசெய்தல் வழிமுறை (EM algorithm) உண்மையான அதிகபட்ச நிகழ்வியல்பு அளகை மதிப்புகளைக் (true maximum likelihood parameter values) கண்டறிவதை உறுதி செய்யாது; மற்றும் உண்மையான அதிகபட்ச நிகழ்வியல்பு அளகை மதிப்பீடுகள் (true maximum likelihood estimates) பெறப்பட்டாலும் பேச்சு அறிதலுக்கு மிகச் சிறந்தவைகளாக இருக்காது. எனவே சரியான தொடக்கச் செயல்முறைகள் மற்றும் அளகை மதிப்புகளின் மீது கட்டுப்பாடுகளைப் பயன்படுத்தல் போன்ற நடைமுறைபடுத்தும் விளக்கங்கள் மிக முக்கியமானதாக இருக்கவியலும்.

பயிற்சியின் நோக்கம் கவனிக்கப்பட்ட தரவுக்கு விளக்கமளிக்க நல்ல மாதிரியைக் கண்டுபிடிப்பது என்பதால் பேச்சு அறிவானின் (speech recognizer) செயல்திறன் பயிற்சித் தரவின் பிரதிநிதித்துவத்தை முற்றிலும் சார்ந்து இருக்கின்றது. இந்தச் சார்பைக் குறைப்பதற்கான சில முறைகள் அடுத்தபகுதியில் கூறப்பட்டுள்ளன. பெரிய பேச்சாளர் தொகையிலிருந்து பெறப்பட்ட தரவைக் கொண்ட பெரிய பேச்சு விரிதரவுகள் மீதான ஒலியியக்க மாதிரிகளின் அளகைகளை மதிப்பீடுசெய்து பேச்சாளர் சுதந்திரம் பெறப்பட்டுள்ளது. உடற்கூறியியல் மற்றும் சமூக வேறுபாடுகளின் காரணமாக ஆண் பேச்சாளர்களுக்கும் பெண் பேச்சாளர்களுக்கும் இடையில் பேச்சு வடிவில் கணிசமான வேறுபாடுகள் உள்ளன. (சராசிரியாகப் பெண்களுக்குப் பேச்சுக் குழலின் நீளம் குறைவாக உள்ளதால் இது அதிக ஒலிச்செறிவு (formant) நிகழ்வெண்களையும் அதிக அடிப்படை நிகழ்வெண்களையும் விளைவிக்கும்; குரல்வளை மடல்கள் (vocal folds) முழுமையாக மூடாததன் காரணமாகப் பெண்களின் குரல் பெரும்பான்மையும் மூச்சிரைப்புடன் இருக்கும்). இதன் காரணமாகப் பேச்சு அறிதல் திறனை மேம்படுத்த ஆண்களுக்கும் பெண்களுக்கும் தனித்தனியான மாதிரிகளைப் பயன்படுத்துவது பொதுவான நடைமுறையாகும்; இது முறைப்படியாகத் தானியங்கு பால் அடையாளங்காணலை வேண்டும்.

3.6.8.3. தழுவல் (adaptation)

இப்பகுதியில் தொடர்ச்சியான அடர்த்தி எச்.எம்.எம். கொண்டு பயன்படுத்தப்படுகின்ற நுட்பங்கள் பற்றி விளக்கப்படும் (Lamel and Gauvain 2003:311-312).

பயிற்சி நிலைமைகளுக்கும் (training conditions) சோதனை நிலைமைகளுக்கும் (test conditions) இடையில் பொருத்தமின்மை (mismatch) இருக்குமானால் பேச்சு அறிவான்களின் (speech recognizers) திறன்கள் குறையும். பொருத்தமின்மையைக் குறைக்க பொருத்தமான நிலைமைகளில் கிடைக்கும் துல்லியத்தைப் பெறுவதற்குப் பல அணுகுமுறைகள் பயன்படுத்தப்படுகின்றன. ஒலியியக்கவியல் சூழல் (acoustic environment), ஒலிவாங்கி (microphone) மற்றும் ஒலிபரப்பு அலைவரிசைகள் (transmission channels), அல்லது ஒரு குறிப்பிட்ட பேசுபவரின் சிறப்பியல்புகள் இவற்றில் உள்ள வேறுபாடுகள் போன்ற பயிற்சி மற்றும் சோதனை கட்டுப்பாடுகளுக்கு இடையே உள்ள பொருத்தமின்மைகளை ஈடுசெய்ய ஒலியியக்க மாதிரித் தழுவலைப் பயன்படுத்த இயலும். இந்த நுட்பங்கள் இரைச்சல் ஈடாக்கம் (noise compensation), அலைவரிசைத் தழுவல் (channel adaptation), பேசுபவர் தழுவல் (speaker adaptation) என முறையே பொதுவாக அழைக்கப்படுகின்றன. பொதுவாக அலைவரிசை வகை (channel type), பின்னணி இரைச்சல் சிறப்பியல்புகள் (background noise characteristics) அல்லது பேசுபவர் இவற்றின் முன்னறிவு இல்லாததன் காரணமாக கண்காணிக்கப்படாத முறையில் (unsupervised mode) சோதனைத் தரவை (test data) மட்டும் பயன்படுத்தி தழுவல் நிறைவேற்றப்படுகின்றது.

பல நேர்வுகளில் பிரதிநிதித்துவத் தரவுகள் குறைந்த அளவிலேயே கிடைப்பதால் குறைந்த அளவிலான தரவை ஈடுசெய்ய வேண்டி ஒலியியக்க மாதிரிப் பயிற்சியில் (acoustic model training) ஒரே கருவிகளைப் பயன்படுத்த இயலும். பிற பெரிய அளவிலான மூலத் தரவின் மீது பயிற்சிசெய்யப்பட்ட மாதிரிகளைத் தழுவச்செய்ய பிரதிநித்துவத் தரவின் சிறிய அளவைப் பயன்படுத்துவது அடிப்படையான கருத்தாகும். பால் சார்ந்த (gender-dependent), பேசுபவர் வரையறுக்கப்பட்ட (speaker-specific), அல்லது செய்பாடு வரையறுக்கப்பட்ட (task-specific) மாதிரிகளை உருவாக்குவது அல்லது செயல்திறனை மேம்படுத்த பேச்சுபவர் தழுவிய பயிற்சியைப் (speaker adaptive training (SAP)) பயன்படுத்துவது என்பன சில எடுத்துக்காட்டான பயன்பாடுகள் ஆகும்.

எச்.எம்.எம்.-இன் அளகைகளைத் தழுவ மூன்று பொதுவாகப் பயன்படுத்தப்படும் திட்டங்களை வேறுபடுத்த இயலும்: பெய்சியன் தழுவல் (Bayesian adaptation), நேரியலான மாற்றங்கள் (linear transformation) அடிப்படையிலான தழுவல், மாதிரி இணைப்பாக்க நுட்பங்கள் (model composition techniques). மாதிரி அளகைகள் மீது நிகழ்தகவுக் கட்டுப்பாடுகளைச் சேர்த்துப் பயிற்சி செயல்முறையில் முன்னறிவை உட்படுத்தும் வழியாகப் பெய்சியன்

கணக்கீட்டை காணவியலும். எதிர்பார்ப்பு உச்சமடையசெய்தல் வழிமுறையைக் (EM algorithm) கொண்டு, ஆனால் அதிகபட்ச விதிவரு முறையிலான (maximum a posteriori (MAP) மீள்கணக்கீட்டு வாய்பாடுகளைப் (re-estimation formulas) பயன்படுத்தி எச்.எம்.எம். அளகைகள் இன்னும் கணக்கிடப்படுகின்றது. இது அதிகபட்ச விதிவரு முறை (MAP) தழுவல் நுட்பத்திற்குக் கொண்டுசெல்கின்றது; இதில் எச்.எம்.எம். அளகைகள் மீதான கட்டுப்பாடுகள் நிலவிலிருக்கின்ற மாதிரியின் அளகைகளின் அடிப்படையில் கணக்கிடப்படுகின்றது. பேசுபவர் சுதந்தரமான ஒலியியக்க மாதிரிகள் பால் சிறப்பீட்டு தரவைப் பயன்படுத்திப் பால் தழுவலுக்குத் தொடக்க மாதிரிகளாகச் செயல்பட இயலும். அடையாளப்படுத்தப்பட்ட பயிற்சித் தரவுகள் கிடைக்கின்ற அதிகபட்ச விதிவரு முறையிலான (maximum a posteriori (MAP) தழுவல் எந்த விருப்படுகின்ற நிலைமையையும் தழுவுவதற்குப் பயன்படுத்த இயலும். நேரியல் மாற்றங்கள் அதிகபட்ச நிகழ்வியல்பு பயிற்சிசெய்யப்பட்டவை மற்றும் எச்.எம்.எம். காசியன் சராசிரிகளுக்குப் பயன்படுத்தப்படுகின்றது; மட்டுமன்றி காசியன் மாறுபாட்டு அளகைகளுக்கும் பயன்படுத்த இயலும். அதிகபட்ச நிகழ்வியல்பு நேரியல் பின்னடைவு நுட்பம் கண்காணிக்கப்படாத தழுவலுக்குப் பொருத்தமானதாகும்; ஏனென்றால் தழுவல் அளகைகள் மிகக்குறைவானவை ஆகும். எம்.எல்.எல்.ஆர். தழுவல் பரிசோதனைத் தரவுக்கும் பயிற்சித் தரவுக்கும் பயன்படுத்த இயலும். பின்னணி இரைச்சலை வெளிப்படையாக மாதிரிப்படுத்தி மற்றும் இந்த மாதிரியைச் சுத்தமான பேச்சு மாதிரியுடன் இணைத்துக் கூட்டத்தக்க இரைச்சலுக்கு ஈடுசெய்ய மாதிரி இணைப்பாக்கம் பெரும்பாலும் பயன்படுத்தப்படுகின்றது. இந்த அணுகுமுறை இரைச்சல் அலைவரிசையை நேரடியாக மாதிரிப்படுத்தும் அனுகூலம் உள்ளது; இது ஒரே சிக்கலுக்கு எம்.எல்.எல்.ஆர். நுட்பத்தால் நிறைவேற்றப்படும் கண்மூடித்தனமான தழுவலுக்கு எதிரானது.

தெரிந்தெடுக்கப்பட்ட தழுவல் முறை பொருத்தமின்மையின் வகைப் பொறுத்தும் கிடைக்கின்ற தழுவல் தரவின் அளவைப் பொறுத்தும் அமையும். ஒலியியக்கத் தொடக்க மாதிரிகளின் தழுவல் புதிய விரிதரவுக்கோ அல்லது பயிற்சித் தரவின் துணைகுழுமத்திற்கோ இருப்பது போன்று தழுவல் தரவு பயிற்சித் தரவின் (பால், கிளைமொழி, பேசுபவர், அல்லது ஒலியியக்க நிலைமை திட்டவட்டம்) பகுதியாக இருக்கலாம் அல்லது பயிற்சித் தரவாக (அதாவது ஒலிபெயர்தல் செய்யவேண்டிய தரவு) இருக்கலாம். தழுவல் தரவுக்கு நோக்கீட்டு ஒலிபெயர்ப்பு கிடைக்கக்கூடும் என்பதால் முந்தைய நேர்வில் கண்காணிக்கப்பட்ட தழுவல் நுட்பங்களைப் பயன்படுத்த இயலும். பிந்தைய நேர்வில் கண்காணிக்கப்படாத நுட்பங்களைப் பயன்படுத்த இயலும்.

3.6.9.சொல்சார் மற்றும் உச்சரிப்பு மாதிரியாக்கம்

ஒலியியக்க நிலை உருப்படுத்ததிற்கும் (acoustic-level representation) பேச்சு அறிவானில் சொற்கோர்வை வெளியீட்டிற்கும் (word sequence output) இடையில் இருக்கும் தொடர்பு அகராதி (lexicon) ஆகும் (Lamel and Gauvain 2003:311-313). சொல் வடிவமைப்பு (lexical design) இரண்டு முக்கியப் பகுதிகளைத் தற்கோள் செய்யும்: சொற்றொகைச் சொற்களின் வரையறைவிளக்கமும் தேர்வும் மற்றும் அறிவானின் அடிப்படை ஒலியியக்க அலகுகளைப் பயன்படுத்தி ஒவ்வொரு உச்சரிப்பு பதிவின் உருப்படுத்தம். அறிதல் திறன் வெளிப்படையாக சொல்சார் உள்ளடக்கப் பரப்புடன் தொடர்புடையது மற்றும் ஒலியியக்க மாதிரிகளின் துல்லியம் ஒவ்வொரு சொல் பதிவுடன் தொடர்புடைய உச்சரிப்பின் உறுதிப்பாட்டுடன் தொடர்புகொண்டிருக்கும்.

அறிதல் சொற்றொகை ஒரு தரப்பட்ட அளவிலான அகராதியின் சொல்சார் உள்ளடக்கப் பரப்பை மிகுதியாக்கப் பெரும்பாலும் தேர்ந்தெடுக்கப்படுகின்றது. சராசரியாகச் சொற்றொகையில் இல்லாத சொல் ஒரு தனிப் பிழையை விடக் கூடுதல் பிழையை விளைவிப்பதால் அறிதல் சொற்றொகையை மதிநுட்பத்துடன் தேர்ந்தெடுப்பது முக்கியமாகும். சொற்பட்டியல் தேர்வு அடுத்த தலைப்பில் விளக்கப்பட்டுள்ளது. தெரிந்தெடுக்கபட்ட தொடக்க அலகுகளைப் (பெருபாலும் ஒலியன்கள் அல்லது ஒலி போன்ற அலகுகளை) பயன்படுத்தி விளக்கப்பட்டுள்ள ஒன்றோ அதற்கு மேலோ உள்ள உச்சரிப்புகள் ஒவ்வொரு சொற்பதிவுடன் தொடர்புபடுத்தப்பட்டுள்ளது. இக்குழுமஅலகுகள் தெளிவாக மொழிச் சுதந்திரமானவை. எடுத்துக்காட்டாகப் பொதுவாகப் பயன்படுத்தப்படும் ஒலிக் கணம்/குழுமம் ஆங்கில மொழிக்கு 45, ஜெர்மன் மொழிக்கு 49 , பிரஞ்சு மொழிக்கு 35, ஸ்பானிஷ் மொழிக்கு 26 ஆகும். உச்சரிப்பு அடிப்படை வடிவங்களை உருவாக்குவதில் பெரும்பாலான அகராதிகள் தரமான உச்சரிப்புகளை உட்படுத்தும் மற்றும் வெளிப்படையாக மாற்றொலிகளை (allophones) உருப்படுத்தம் செய்யாது. இந்த உருப்படுத்தங்கள் பெரும்பாலான மாற்றொலி பேதங்களாக விதிகளால் வருவதுரைக்க இயலும் மற்றும் அவற்றின் பயன்பாடு விருப்பாகும். மிகமுக்கியமாக ஒரு தரப்பட்ட ஒலியனின் வேறுபட்ட மாற்றொலிகளுக்கு இடையே பெரும்பாலும் ஒரு தொடர்ச்சி காணப்படும்; எது ஒரு தரப்பட்ட கூற்றில் வருகின்றது என்பது அகவயமானதாகும். ஒலியன்சார் உருப்படுத்தத்தைப் பயன்படுத்தி எந்த உறுதியான தீர்மானமும் திணிக்கப்படுவதில்லை; பயிற்சித் தரவில் கவனிக்கப்பட்ட மாற்றுருக்களை உருப்படுத்தம் செய்வது ஒலியியக்க மாதிரிகளிடம் விடப்பட்டுள்ளது. உச்சரிப்பு அகராதிகள் பெரும்பாலும் (அல்லது பகுதி அளவிலாவது) மனிதமுயற்சியால உருவாக்கப்படும் போது தானியக்கமாகச் சொல் உச்சரிப்புகளைக் கற்பதற்கும் உருவாக்குவதற்கும் பல அணுகுமுறைகள் ஆயப்பட்டுள்ளன.

மாற்று உச்சரிப்பு மாற்றுருக்கள் கவனிக்கப்பட்டுள்ள பலவகைச் சொற்கள் உள்ளன; இந்த மாற்றுருக்கள் ஈருயிருடன் உச்சரிக்கப்படும் (/ɑᶦ/ அல்லது ஸ்வா /ə/) *ization* என்ற பின்னொட்டில் காணப்படுவது போன்று மாற்றொலி வேறுபாடுகள் அல்ல. excuse, record, produce போன்ற வேறுபட்ட சொல்வகைபாடுகளைச் (வினையை அல்லது பெயரை) சார்ந்த ஒப்பெழுத்துச் சொற்களுக்கும் (homographs) (ஒரே எழுத்துக்களை உடைய ஆனால் வெவ்வேறு உச்சரிப்புகளைக் கொண்ட சொற்கள்) மாற்று உச்சரிப்புகள் தேவை. interest, company போன்ற மூன்று அசைகளைக் கொண்ட சொற்கள் பெரும்பான்மையும் இரண்டு அசைகளாக உச்சரிக்கப்படுகின்றன. எடுத்துக்காட்டாக interest என்ற சொல் வேறுபட்ட பேசுபவர்களால் வேறுபட்டு உச்சரிக்கபடுகின்றன. ஒரு நேர்வில் interest /ɪntrɪst/ என்பது t நீக்கப்பட்டும் n மூக்கு வருடொலியாகவும் /ɪnɜ˞ɪs/ என்றும் மற்றொரு நேர்வில் இரு அசைகளாக விருப்பு உயிர் இன்றி tr மயக்கமாக /ɪntrɪs/ என்றும் உச்சரிக்கப்படுகின்றது. தேவைப்பட்ட மாற்றுருக்களை உள்ளடக்கிய உச்சரிப்பு அகராதியுடன் கூடிய உகந்த வரிசையாக்கம் /intrist/, /intris/, /ɪnɜ˞ɪs/ என்பவைகளைக் கொண்டிருக்கவேண்டும். சரியான வரிசையாக்கம் கூடுதல் துல்லியமான ஒலியியக்க ஒலி மாதிரியில் விளையும். கவனமான சொல்சார் வடிவமைப்பு பேச்சு அறிதல் திறனை மேம்படுத்தும்.

விரைவாகப் பேசுபவரின் அல்லது தளர்ந்து பேசும் நடைகள் உள்ள பேசுபவர்களின் பேச்சு வடிவில் அழுத்தமில்லாத அசைகள், குறிப்பாக அழுத்தமில்லாத அசைகளின் வரிசைகள் கொண்ட நீண்ட சொற்களில், தெளிவற்று உச்சரிக்கப்படுவது (அல்லது உச்சரிக்காமல் விடப்படுவது) பொதுவான நேர்வாகும். நீண்ட சொற்கள் எடுத்துக்காட்டாக நன்றாக அறியப்பட்டாலும் பெரும்பாலும் அடுத்துவரும் செயல்பாட்டுச் சொல் நீக்கப்படுகின்றது. இவ்வகையிலான தவறுகளை குறைக்க positioning (/pəzɪʃənɨŋ/ அல்லது(/pəzɪʃənɨŋ/) போன்ற நீண்ட சொற்களுக்கு மாற்று உச்சரிப்புகளை ஸ்வா நீக்கம் அல்லது அழுத்தமில்லாத அசைகளில் அசை மெய்கள் இவற்றை அனுமதித்து அகராதியில் உட்படுத்தப்படவேண்டும். ‘dija’ என்று உச்சரிக்கப்படும் ‘did you’ அல்லது ‘gonna’ என்று உச்சரிக்கப்படும் ‘going to’ போன்ற பொதுவான சொல் கோர்வைகளுக்குக் குறைக்கப்பட வடிவுகளை உருப்படுத்தம் செய்யும் வழியாகக் கூட்டுச் சொற்கள் பயன்படுத்தப்படுகின்றன. மாறாக அம்மாதிரியான சரளமான பேச்சு விளைவுகளை ஒலியனியல் விதிகளைப் பயன்படுத்தி மாதிரியாக்கம் செய்ய இயலும். இந்த ஒலியனியல் விதிகளுக்குப் பின்னால் வரும் கொள்கைகள் இம்மாதிரியான வேறுபாடுகளைக் கணக்கில் எடுக்க அனுமதிக்க இயலும் ஒலிக்கோர்வைகளை மாற்றுவதற்கு ஆகும். இவ்விதிகள் பயிற்சியின் போதும் அறிதலின் போதும் விருப்பாகப் பயன்படுத்தப்படுகின்றது. ஒலியன் விதிகள் தவறான ஒலிபெயர்ப்பால் மாசுபடுத்தப்படுவது குறைவு ஆகையால் பயிற்சியின் போது ஒலியன்

விதிகளைப் பயன்படுத்துவது சிறந்த மாதிரிகளில் விளையும். அறிதலின் போது அவற்றைப் பயன்படுத்துவது தவறுகளின் எண்ணிக்கையைக் குறைக்கும் (Lamel and Gauvain 2003:311-313).

3.6.10. தமிழில் பேச்சிலிருந்து உரையாக்கம்

பேச்சிலிருந்து உரை (Speech to text (STT) அதாவது பேச்சை உரையாக மாற்றுவது இயற்கைமொழி ஆய்வின் ஒரு சவாலான பணியாகும். இதைப்பற்றி விரிவாக முன்னர் பார்த்தோம். தமிழில் பேச்சிலிருந்து உரை உருவாக்க முயற்சிகள் பற்றி இத்தலைப்பில் விளக்கப்படுகின்றது.

3.6.10.1. தமிழில் பேச்சிலிருந்து உரை உருவாக்க முயற்சிகள்

யெக்னநாராயணா (Nayeemulla Khan and Yegnanarayana 2001.), ஹேமா ஏ.மூர்த்தி (Lakshmi A. and Hema A. Murthy. 2006), ஏ.ஜி.ராமகிருஷ்ணன் மற்றும் டி.நாகராஜன் ஆகியோர் பேச்சை உரையாக மாற்றுவதற்கான அடித்தளத்தை அமைத்துள்ளனர். ஏ.ஜி.ராமகிருஹ்னன் ஒரு பெரிய சொற்றொகையை உள்ளடக்கிய தமிழ் வெளிப்பாடுகளை அறிய நரம்பியல் பின்னல் மற்றும் முற்றுநிலை மாறிகளைப் பயன்படுத்தி ஒரு தானியங்கி பேச்சு அறிதல் முறையை உருவாக்கினார். டி. நாகராஜன் DST-TIDE- நிதி உதவியுடன் "பெருமூளை வாதம் கொண்ட பேச்சாளர்களுக்கான பேச்சு-உள்ளீட்டு பேச்சு-வெளியீட்டு கருத்துப்பரிமாற்ற தொடர்பு உதவிக்கருவி' (Speech-Input Speech-Output Communicatin Aid (SISOCA) for speakers of Cerebral Palasy) என்ற மொன்பொருளை உருவாக்கினார். 'பேச்சிலிருந்து உரை'க்காக கூகிள் ஆட்மொப் (Google AdMob) ஒரு மென்பொருளைக் கொண்டுள்ளது. இது ஒலிவாங்கி மூலம் பேசப்படும் குரலை அடையாளம் கண்டு தமிழ் உரையைத் தானாகத் தட்டச்சு செய்யும். இந்தத் தட்டச்சு செய்த உரையை ஒருவர் சேமித்து எங்கும் பயன்படுத்தலாம். இந்த மென்பொருள் Google Chrome (பதிப்பு 25 அல்லது அதற்கு மேற்பட்டது) உலாவியில் (browser) மட்டுமே செயல்படும். கூகிள்பிளே தமிழ் பேசிலிருந்து உரை மாற்றியை வழங்குகிறது. அழகி ஆண்ட்ராய்டு பயன்பாடு-குரல் உள்ளீடு என்பது 100+ மொழிகளுக்கான பேசிலிருந்து உரை மாறியாகும். அழகி ஆண்ட்ராய்டு பயன்பாட்டைப் பயன்படுத்தும் போது, தட்டச்சு செய்வதைத் தவிர, ஒருவர் தமிழ் உட்பட பல மொழிகளில் பேசலாம் மற்றும் ஒருவரின் உரைகளை உருவாக்கலாம். கோயம்புத்தூர் கொங்கு பொறியியல் கல்லூரியின் தங்கராஜன் தமிழிற்கான பேசிலிருந்து உரை அமைப்பின் வளர்ச்சிக்கும் பங்களித்துள்ளார் (Thankarajan et al 2008 a, 2008 b). அகிலா மற்றும் பிறர் (Akila

and et al 2013, 2015) தமிழில் பேச்சிலிருந்து உரை உருவாக்கத்திற்கான முயற்சிகள் மேற்கொண்டுள்ளனர்.

3.6.10.2. பேச்சுத்தரவுகள்

பேச்சிலிருந்து உரையாக்கத்திற்குப் பேச்சுத்தரவு முக்கியமானதாகும். மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் தமிழ்ப் பேச்சுத்தரவை உருவாக்கியுள்ளது. இதன் விவரம் கீழே தரப்பட்டுள்ளது.

மொத்தம் பேசுபவர்கள் = 453

மொத்த மணிகள் = 213:37:27

குறிப்பு: பேச்சுத் தரவின் மொத்த மணிகள் தரமான பரிசோதனை முழுமையடையும் போது மாறும்.

பேச்சுத்தரவு அசைகளாகப் பிரிக்கப்படவேண்டும். அசைகளும் ஒலிகளும்/ஒலியன்களும் அடையாளப் படுத்தவேண்டும். இந்த அடையாளபடுத்தப்பட்ட பேச்சுத்தரவைப் பயிற்சி கட்டதிலும் பரிசோதனைக் கட்டத்திலும் பயன்படுத்தலாம்.

3.6.10.3.தமிழுக்கான அசை அடிப்படையிலான தொடர்ச்சியான பேச்சு அறிவான் (A Syllable Based Continuous Speech Recognizer for Tamil)

லக்ஷ்மி மற்றும் ஹேம எ. மூர்த்தி என்போரின் (Lakshmi A. and Hema A. Murthy. 2006) A Syllable based continuous speech recognizer for Tamil (தமிழுக்கான அசை அடிப்படையிலான தொடர்ச்சியான பேச்சு அறிவான்) குறிப்பிட தகுந்த ஆய்வாகும். லக்ஷ்மி மற்றும் ஹேம எ. மூர்த்தி என்போர் தமது கட்டுரையில் அடையாளப்படுத்தப்படாத ஒலிபெயர்ப்பு செய்யப்பட்ட பயிற்சித் தரவு இருக்கும் போது அசையை அடிப்படையாகக் கொண்ட பேசு அறிவானை உருவாக்கும் ஒரு புதிய நுட்பத்தை முன்வைக்கிறனர். அவர்கள் பேச்சையும் மற்றும் பொருத்தமுள்ள உரையையும் ஒப்பிடத்தக்க அசைபோன்ற அலகுகளாகப் கூறிடவும் இரண்டு வெவ்வேறு கூறிடும் வழிமுறைகளை (two different segmentation algorithms) முன்வைக்கின்றனர். பேச்சுத் தரவிலிருந்து துல்லியமாக அசைபோன்ற அலகுகளைக் கூறிட ஒரு குழு தாமதத்தை அடிப்படையாகக் கொண்ட இரு நிலை கூறிடல் வழிமுறையை (group delay based two level segmentation algorithm) முன்மொழிகின்றனர். ஒரு விதி அடிப்படையிலான உரை கூறிடல் வழிமுறை பேச்சுடன் தொடர்புடைய உரையைத் தானியக்கமாக அசை அலகுகளாக அடையாளப்படுத்தப் பயன்படுத்தப்படுகின்றது. அடையாளப்படுத்தப்பட பேச்சிலிருந்து எடுத்துக்காட்டுகளைச் சேகரிப்பதன் மூலம் எல்லாத் தனித்துவமான அசைகளுக்கும் பன்முக சட்ட

அளவையும் (multiple frame size (MFS)) மற்றும் பன்முகச் சட்ட வீதத்தையும் (multiple frame rate (MFR)) பயன்படுத்தி தொடர்பற்ற பாணி அசை மாதிரிகள் (isolated style syllable models) உருவாக்கப்பட்டுள்ளன. தமிழ் மொழியில் நிகழ்த்தப்பட்ட சோதனைகள் பேச்சறியும் செயல்திறன் கைமுறையைப் பயன்படுத்தி கூறிடப்பட்ட பயிற்சித் தரவைப் பயன்படுத்தி உருவாக்கப்பட்ட அறிவானுடன் ஒப்பிடத்தக்கது என்று காட்டுகின்றது. இந்தச் சோதனைகள் பேச்சுத் தரவின் வாக்கிய நிலை ஒலிபெயராக்கம் கிடைத்தால் பேசு அறியும் ஒழுங்குமுறையை உருவாக்கும் செலவைக் குறைக்க இயலும் என்று பரிந்துரைக்கின்றன.

தொடர்ச்சியான பேச்சு அறியும் ஒழுங்குமுறையை உருவாக்குவதில் இன்றைய சவால் தேவையான அடையாளப்படுத்தப்பட்ட ஒலிபெயர்க்கப்பட்ட பயிற்சித் தரவைப் பெறுவதற்கான செலவு ஆகும்; இது மனித சக்தி மற்றும் நேரம் என்ற இரண்டின் அடிப்படையில் ஒரு விலையுயர்ந்த செயல்முறையாகும். ஒலிபெயர்ப்பு செய்யப்படாத தரவைப் பயன்படுத்தி கண்காணிக்கபடாத பயிற்சி நுட்பங்கள் அடிப்படையில் பல ஆராய்ச்சிகள் மேற்கொள்ளப்பட்டுள்ளன. இந்த உத்திகள் பெருமளவில் நம்பகமான பேச்சு அறிவான் (trusted recognizer), பெரிய அளவிலான ஒலியியக்கவியல்சார் தரவு (acoustic data) மற்றும் ஒரு நல்ல மொழி மாதிரியைப் (language model) பெரிதும் நம்பியுள்ளது. அசைகள் ஒலியியக்க அடிப்படையிலும் கேட்பு அடிப்படையிலும் நிலையான அலகுகள் என்று அறியப்படுவதால் கடந்த பத்தாண்டுகளாக அசை அடிப்படையிலான பேச்சு அறிதல் ஒழுங்கமைப்புகளின் முக்கியத்துவம் உணரப்பட்டுள்ளன. அசை அடிப்படையிலான பேச்சு அறிவானுக்குப் பயிற்சி அளிக்கக் கண்காணிக்கப்படாத அதிகரிப்புசார் நுட்பங்கள் (unsupervised incremental clustering techniques) முயற்சிக்கப்பட்டன (Sarada, Nagarajan Hemalatha and Hema A. Murthy 2004). இந்த நுட்பங்கள் அதிகரிப்பு அடிப்படையில் உருவாக்கப்பட்ட பேச்சு அறிவானைப் (incrementally built recognizer) பயன்படுத்தி அசைக் கொத்துக்களை/திரள்களைத் தானியக்கமாக உருவாக்கப்பட அனுமதிக்கின்றது மற்றும் பின்னர் கொத்துக்கள்/திரள்கள் கைமுறையாக புலக்குறிப்பு செய்யப்படுகின்றன. இந்த அணுகுமுறையின் குறைபாடு வேறுபட்ட ஒன்றுபோல் ஒலிக்கும் அசைகள் பெரும்பாலும் ஒன்றாகக் கொத்தாகி கொத்துக்களைப் புலக்குறிப்பு செய்வதில் சிக்களை ஏற்படுத்துகின்றன. குறைவான துல்லியமாகக் கொத்தாகத் தரவு (clustered data) மற்றும் கொத்துக்களைப் புலக்குறிப்பு செய்வதற்கான மனித உழைப்பு என்பன இந்த நுட்பங்கள் குறைவாக விரும்பத்தக்கவையாகச் செய்கின்றன.

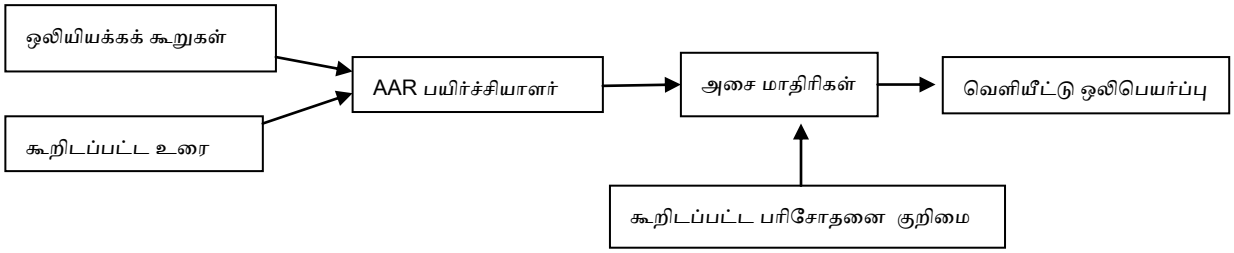
தற்போது சாரதா மற்றும் கூட்டாளிகள் (Sarada, Nagarajan Hemalatha and Hema A. Murthy 2004) பேச்சு குறிகைக்கு (speech signal) வாக்கிய நிலை ஒலிபெயர்ப்புகள் இருக்கின்றன என்று அனுமானிக்கின்றனர். அசை நிலை கூறிடல் வழிமுறை ஒலியியக்க குறிகை மீது பிரயோகிக்கப்படுகின்றது. இந்த அலகுகள் தானியக்க அடையாளப்படுத்தலை உருவாக்கப் பொருத்தப்படுகின்றன.

முந்தைய செயல்பாடு (Kamashi Prasad et al 2004) குழு தாமத அடிப்படையிலான கூறிடல் வழிமுறை கூறு எல்லைகளைச் சரியாக அடையாளம் காட்டினாலும் துல்லியமாக, 30% அசை எல்லைகள் தவறப்பட்டன. நாகராஜன் கூட்டாளிகள் (Nagarajan and Hema A Murthy 2004) எல்லைக் கண்டறிதலின் வழிமுறை மாற்றி 10% முன்னேற்றத்திற்கு வழிவகுத்தனர். எல்லை கண்டறிதல் வழிமுறைகளின் வரம்பு எல்லைகளின் பேச்சு குறிகைப் பண்புகளின் துல்லியம் குறிப்பாகக் கூற்றைத் தாண்டி மாறி அசை விகிதம் சார்ந்தது என்ற உண்மையாகும். சாரதா மற்றும் கூட்டாளிகள் இரண்டு நிலை கூறிடல் வழிமுறையை முன்மொழிகின்றனர்; இது தேவைப்பட்டால் அசை காலத் தகவல்களை அலகுகளை மறு கூறிடலுக்கு பயன்படுத்துகிறது. சார்ல்ஸ் மற்றும் பிறர் தமிழ் பேச்சு அறிவான் உருவாக்க முயற்சித்ததாக அவர்களது கட்டுரையிலிருந்து தெரியவருகின்றது (Charales A.P.H. and Devaraj G. 2004).

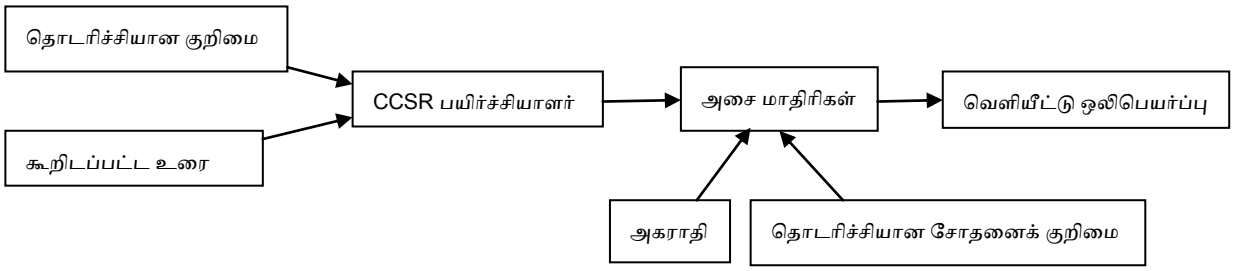
உரை கூறிடல் என்பது மொழியிலிருந்து பெறப்பட்ட மொழியியல் விதிகளை அடிப்படையாகக் கொண்டது. எந்த அசை அடிப்படையிலான மொழியையும் பொதுவான விதிகளைப் பயன்படுத்தி அசைகளாகப் பிரிக்கலாம். உரைக் கூறுகளைப் பேச்சு அலகுகளுக்கு சமமாக சரியாக உருவாக்கச் சில மொழிச் சிறப்பு விதிகள் பயனுள்ளதாக இருக்கும். உரை கூறிடப்பட்ட பிறகு கூறிடப்பட்ட அலைவடிவமும் கூறிடப்பட்ட உரையும் பொருத்தப்படும். ஒவ்வொரு தனித்துவமான அசையின் பன்முக மெய்ப்படுத்தங்கள் பிரித்தெடுக்கப்படுகின்றன மற்றும் HMMகள் ஒவ்வொரு அசைக்கும் பயிற்சி அளிக்கப்படுகின்றன. ஒலியியக்க மாதிரிகள் பன்முகச் சட்ட அளவுகள் (multiple frame sizes (MFS)) மற்றும் பன்முகச் சட்ட விகிதங்கள் (multiple frame rates (MFR)) பண்புக்கூறு பிரித்தெடுத்தல் அடிப்படையிலான நுட்பம் பயன்படுத்தி உருவாக்கப்பட்டுள்ளன. இது காலப்போக்கில் நிறமாலை மாறுபாடுகளுக்கும் பயிற்சித் தரவின் குறைவுக்கும் காரணமாகிறது. தானியக்கமாக அடையாளப்படுத்தப்பட்ட பேசு அறிவானின் செயல்திறன் (automatically annotated recognizer (AAR)) வழக்கமான HMM அடிப்படையிலான தொடர்ச்சியான பேச்சு அறிவானுடன் (conventional HMM based continuous speech recognizer CCSR) ஒப்பிடப்படுகின்றது. AAR பேச்சு அறிவான் அசைக்காகக்

கூறிடப்பட்ட உரையையும் கூறு எல்லைத் தகவல்கொண்ட தொடர்ச்சியான பேச்சையும் தொடர்ச்சியான பேச்சு அறிவானுக்காக அசை மாதிரிகளை உருவாக்கப் பயன்படுத்துகிறது; சி.சி.எஸ்.ஆர். எல்லைகளைத் தானியக்கமாகக் கண்டறிந்து வரிசைப்படுத்துவதன் மூலம் தொடர்ச்சியான பேச்சை ஒலிபெயர்ர்பு செய்கின்றது.

AAR அடிப்படையிலான ஒழுங்கமைப்பு



CCSR அடிப்படையிலான ஒழுங்கமைப்பு



3.7. சுருக்கவுரை

கணினி ஒலியனியல் என்ற தலைப்பில் அமையும் இவ்வியல் மூன்று முக்கிய பகுதிகளாகப் பிரிக்கப்பட்டு தகவல்கள் தரப்பட்டுள்ளன. முதல் பகுதியில் ஒலியியல், ஒலிப்பியல், ஒலியியக்கவியல் குறித்தும் இரண்டாம் பகுதியில் உரையிலிருந்து பேச்சாய்வு குறித்தும் மூன்றாம் பகுதியில் பேச்சிலிருந்து உரையாக்கம் குறித்தும் விளக்கம் தரப்பட்டுள்ளன.

தொடக்கத்தில் ஒலியியல் ஆய்வு, ஒலிப்பியல் ஆய்வு, ஒலியியக்கவியல் ஆய்வு குறித்த பொதுவான செய்திகள் தரப்பட்டுள்ளன. தமிழ் ஒலிகள் ஒலிப்பியல் அடிப்படையில் பகுத்தாயப்பட்டு விளக்கம் தரப்பட்டுள்ளன. தமிழ் உயிரொலிகள் மற்றும் மெய்யொலிகளின் ஒலிப்பியல் அடிப்படையிலான விளக்கங்கள் தரப்பட்டுள்ளன. தமிழ் ஒலியியக்கவியல் குறித்து

பல ஆய்வுகள் மேற்கொள்ளப்பட்டு வெளிவந்துள்ளன. ஒரு எடுத்துக்காட்டான ஆய்வு அடிப்படையில் தமிழ் ஒலிகள் ஒலியிக்கவியல் அடிப்படையில் ஆயப்பட்டு விளக்கம் தரப்பட்டுள்ளன. தமிழ் உயிரொலிகளின் ஒலிச்சொறிவு அமைப்பு பற்றி விரிவாக விளக்கப்பட்டுள்ளது. தமிழ் மெய்யொலிகளும் ஒலியியக்கவியல் அடிப்படையில் ஆயப்பட்டு விளக்கப்பட்டுள்ளன. தமிழ் உயிரொலிகள், மெய்யொலிகள் இவற்றின் நிறமாலை நிழற்படங்கள்/ ஒலிவர்ணனைப் படங்கள் தரப்பட்டுள்ளன. நிறமாலை நிழற்படங்களை/ஒலிவர்ணனைப் படங்களை ஒலிகளின் பல பரிமாணங்கள் அடிப்படையில் தர இயலும் என்றாலும் பக்கங்களின் எண்ணிக்கை கருதி எடுத்துக்காட்டக ஒரு சில ஒலிவர்ணனைப் படங்களே இங்கு தரப்பட்டுள்ளன.

இரண்டாவது பகுதி உரையைப் பேச்சாக மாற்றும் செயல்பாடு குறித்ததாகும். தொடக்கத்தில் உரையிலிருந்து பேச்சாக்கம் அறிமுகம் செய்யப்படுகின்றது. அடுத்ததாகப் பேச்சொலிகளின் ஒலிபெயர்ப்பு, ஒலியன் விதிகளும் முற்றுநிலை மாறிகளும், உரையிலிருந்து பேச்சுக்குவேண்டி உரையை ஒலியன்களுக்குப் பொருத்துதல், உரையிலிருந்து பேச்சில் மீக்கூறு என்பன பற்றி விளக்கப்பட்டுள்ளன. மேலும் பேச்சுக் கூட்டிணைப்பாக்கத்தின் வரலாறு, பேச்சுக் கூட்டிணைப்பாக்க அணுகுமுறைகள், உரை-பேச்சு ஒழுங்குமுறைகளின் பயன்பாடுகள் இவை பற்றியும் விரிவாக விளக்கப்பட்டுள்ளன. இதன் பின்னர் தமிழில் மேற்கொள்ளப்பட்ட உரையிலிருந்து பேச்சாக்க முயற்சிகள் குறித்தும் விளக்கம் தரப்பட்டுள்ளது.

மூன்றாவது பகுதி பேச்சிலிருந்து உரையாக்கம் அதாவது பேச்சு புரிதல் குறித்ததாகும். பேச்சை உரையாக மாற்றும் செயல்பாட்டிற்கான கோட்பாடுகள் மற்றும் கொள்கைகள் குறித்து முதலில் விளக்கப்பட்டுள்ளன. பின்னர் தமிழில் பேச்சிலிருந்து உரையாக்கத்திற்கு மேற்கொள்ளப்பட்டு வருகின்ற ஆய்வுகள் குறித்தும் விளக்கம் தரப்பட்டுள்ளன.

இயல் 4

கணினிசார் உருபனியல்

4.1. அறிமுகம்

சொற்களைக் கணியால் பகுத்து அதன் பொருள்தரும் கூறுகளை (உருபன்களை) ஆய்வதும் அத்தகைய பொருள்தரும் சிறு கூறுகளைக்கொண்டு சொற்களை ஆக்குவதும் கணினி உருபனியல் ஆய்வாகும். முதலாவது உருபனியல் பகுப்பாய்வு (morphological analysis) என்றும் இரண்டாவது உருபனியல் ஆக்கம் (morphological generation) என்றும் அறியப்படும். சொற்களை எல்லாம் அகராதில் தந்து அதன் பொருள்களைத் தருவது சாத்தியமானதல்ல மட்டுமன்றி சிக்கமானதல்ல. ஆங்கிலம் போன்ற எளிய உருபனியல் அமைப்பு கொண்ட மொழிகளுக்கு இது ஒருவேளை சாத்தியப்படலாம்; ஆனால் தமிழ் போன்ற கலவைத்தன்மையான உருபனியல் அமைப்பு கொண்ட மொழிகளுக்கு இது சாத்தியமல்ல. பின்வரும் செயல்பாடுகளுக்குச் கணிசார் உருபனியல் பகுப்பாவும் உருபனியல் உருவாக்கமும் முக்கியமானதாகும்:

- இயற்கைமொழிப் பயன்பாடுகள் (natural languge applications)
    - பகுப்பாய்வு (parsing)
    - உரை/பனுவல் ஆக்கம் (text generation)
    - இயந்திர மொழிபெயர்ப்பு (translation)
    - அகராதி உருவாக்கம் (dictionary/lexicon making)
    - சொல்லனாக்கம் (lemmatization)
    - உருபனியல் ஆய்வு (morphological analysis)
    - உருபனியல் ஆக்கம் (morphological] generation)
    - அடி உருபன் அல்லது வேர் உருபன்களை ஆய்தல் (stemming)
- பேச்சுப் பயன்பாடுகள்
    - உரையிலிருந்து பேச்சு ஒழுங்கமைப்புகள்
    - பேச்சு அறிதல்
- சொல் ஆய்வுப் பயன்பாடுகள்
    - எழுத்துப்பிழை திருத்தி
    - உரை உள்ளீடு
- ஆவண மீட்பு

4.2. உருபனியல் அறிமுகம்

ஒரு மொழியின் உருபன்களைக் கண்டு பிடித்து அவற்றின் பட்டியலை உருவாக்குவதும் உருபுகளை உருபன்களின் கீழ் மாற்றுருபுகளாக அடையாளங் கண்டு பட்டியலிடுவதும் உருபனியலின் நோக்கமாகும். மொழி அமைப்பின் இரண்டாவது மட்டமாக வருவது உருபனியல் ஆகும். சொற்களின் அக அமைப்பைப் பற்றிக் கூறுவது உருபனியலாகும். உருபனியல் ஆய்ஞர் சொல் வடிவுகளை அவற்றின் ஆக்கக்கூறுகளாகப் (உருபன்களாகப்) பிரித்து அவை ஒவ்வொன்றின் அமைப்பொழுங்கைக் கூற முயல்வர். சொல்வகைப்பாட்டிற்கு ஏற்ப உருபன்களின் இணைப்பு முறைகளும் வருகை முறைகளும் வேறுபடும். ஒருவரது பேச்சு ஒலியன்களின் அடுக்காக மட்டுமே அமையுமானால் அவர் பேசுவதும் அப்பேச்சினை மற்றவர்கள் கேட்பதும் வீண்வேலையாகும். பேசுவதாலும் கேட்பதாலும் கருத்துப்பரிமாற்றம் நிகழ்வதனாலேயே பேச்சினை ஆராய்வது தேவையாகின்றது. கருத்துக்களைப் பரிமாற முடிகிறது என்றால், அதற்குக் காரணம் உருபனமைப்பு ஆகும். வேறுவகையில் சொன்னால் உருபன்களே பொருளை வெளிப்படுத்துகின்றன. பொருளுக்கும் ஒலியனுக்கும் நேரடித் தொடர்பில்லை. ஓர் ஒலியன் தனித்தோ வேறு ஒலியன்களோடு சேர்ந்தோ ஓர் ஒலியக்கூறு ஆகின்றது. இந்த ஒலியக்கூற்றையே உருபன் என்கின்றோம். இது பொருளுடையது. சான்றாகத் தமிழில் ஆ என்னும் ஒலியனுக்குப் பொருள் இல்லை. அதுவே தனித்து ஒரு உருபனாய் (ஓரெழுத்து ஒரு மொழியாய்) வரும்போது இலக்கியத்தில் 'பசு' என்பதைக் குறிக்கின்றது. இவ்வாறே ஆ என்பதும் ண் என்பதும் சேர்ந்து *ஆண்* என்னும் உருபனாக வரும்போதுதான் 'ஆண் தன்மையுடையவன்' என்னும் பொருள் தோன்றுகின்றது. இவ்வாறு பொருள் தருகின்ற மிகச்சிறிய ஒலியக்கூறே உருபன் எனப்படுகின்றது. இங்கு 'மிகச்சிறிய' என்பது 'மேலும் பிரிக்கவியலாத' என்னும் கருத்தில் வழங்கப்படுகின்றது. *நாய்கள்* என்னும் சொல்லை *நாய்* + *கள்* என்று பிரித்து பொருள் கொள்ளலாம். ஆனால் *நாய்* என்பதையோ *கள்* என்பதையோ மேலும் பரித்துப் பொருள் கொள்ளவியலாது. எனவே *நாய், கள்* ஆகிய இரண்டும் உருபன்களாகும்.

மேலே கூறிய எடுத்துக்காட்டே உருபனியலின் மற்றொரு பயன்பாட்டையும் விளக்குகின்றது. *நாய்* என்பது ஒரு விலங்கினையக் குறிக்கின்றது. *கள்* என்பது பன்மையைக் குறிக்கின்றது. அதாவது *கள்* என்ற சொல்லுக்குச் சொற்பொருள் (lexical meaning) இல்லை, இலக்கணப் பொருள்தான் (grammatical meaning) இருக்கின்றது. வேறுவகையில் சொன்னால் உருபன்கள் தான் இலக்கண உறவுகளையும் வெளிப்படுத்துகின்றன. இவ்வாறு ஒரு புறம் ஒலியனியலுக்கும் பொருண்மையியலுக்கும் பாலமாகவும் மறுபுறம் ஒலியனியலுக்கும்

இலக்கணத்திற்கும் பாலமாகவும் உருபனியல் அமைகின்றது. சொற்களின் அக அமைப்பைப் பற்றிக் கூறுவது உருபனியலாகும். உருபனியல் ஆய்ஞர் சொல்வடிவுகளை அவற்றின் ஆக்கக் கூறுகளாகப் பிரித்து அவை ஒவ்வொன்றின் வருகை முறையைக் கூற முயல்வர்.

திரிபும் ஆக்கமும் (Inflection and derivation)

சொல் திரிபு (Inflection)

ஓர் அடிச் சொல் பல்வேறு உருபனியல் அமைப்புகளில் இடம் பெற்று அதனோடு பல்வேறு உருபன்கள் சேர்ந்து வரும் நிலையில் அதன் மூலம் பெறப்பட்ட இறுதி வடிவமானது அது எந்த வகை அடிச் சொல்லிலிருந்து பெறப்பட்டதோ அவ்வடிச் சொல் சார்ந்துள்ள இலக்கணப் பெரும் பிரிவைச் சார்ந்தே எந்தவித மாற்றமும் பெறாமல் அமையும். பின்வரும் எடுத்துக்காட்டில் *படி* என்பது வினையாகச் செயல்படுவதால் அதனுடன் வினைக்குரிய பின்னொட்டுக்கள் சேர்க்கப்பட்டும். அத்தகைய ஒட்டுக்கள் திரிபு ஒட்டுட்கள் எனப்படும்; இச்சொல்லாக்கம் சொல் திரிபு அல்லது திரிபு எனக்கூறப்படும்: படி, படி + த்த் + உ > படித்து, படி +த்த் +அ > படித்த, படி +க்க் +ஆ+மல் > படிக்காமல், படி+த்த்+ஆன் > படித்தான்.

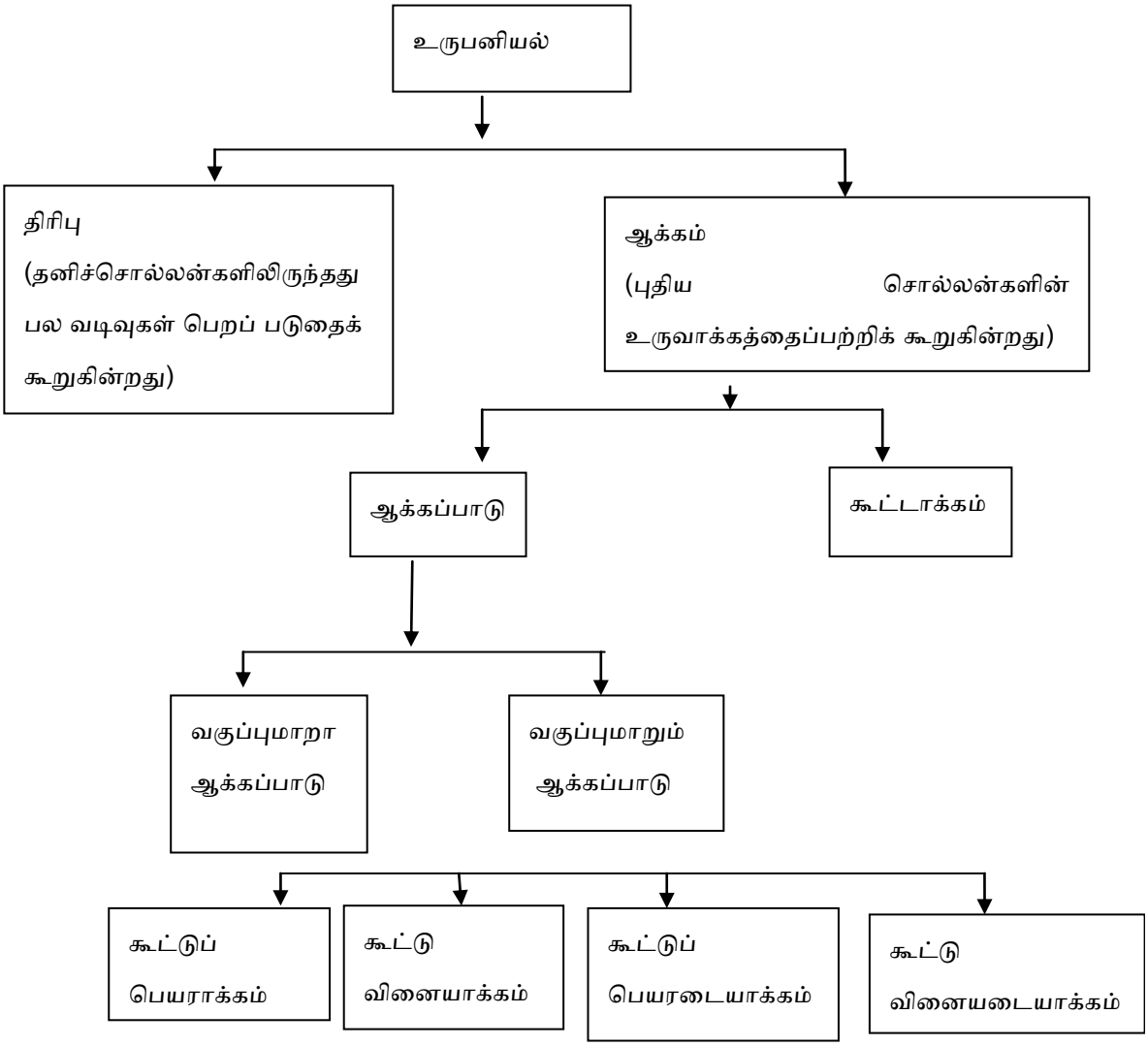
சொல்லாக்கம் (Derivation)

சொற்களை புதிதாக உருவாக்குதலைத் தான் ஆக்கம் என்கிறோம். ஓர் அடிச் சொல் சில உருபனியல் அமைப்புகளில் இடம் பெற்று அதனோடு சில உருபன்கள் சேர்ந்து வரும் நிலையில் பெறப்படும் இறுதி வடிவமானது, அது எந்த வகை அடிச் சொல்லிலிருந்து பெறப்பட்டதோ அதிலிருந்து மாறுபட்டு வேறு இலக்கணப் பெரும் பிரிவைச் சார்ந்ததாக அமையும். இத்தகைய அமைப்பில் முழு மாற்றம் காணப்படுகிறது. அதாவது அடிச் சொல் ஓர் இலக்கணப் பிரிவைச் சார்ந்தாகவும் பெறப்படும் உருபனியல் அமைப்பானது மற்றொரு வகை இலக்கணப் பிரிவைச் சார்ந்ததாகவும் அமைகிறது. எடுத்துக்காட்டாக, தமிழில், வினை + பெயராக்க ஒட்டு > பெயர் என ஆக்கம் பெறும்: பார் + வை > பார்வை, வாழ் + க்கை > வாழ்க்கை, படி + ப்பு > படிப்பு. எனவே வை, க்கை, ப்பு என்பன சொல்லாக்க/ஆக்க ஒட்டுகள் எனப்படும். இச்செயல்பாடு சொல்லாக்கம் எனப்படும்.

உருபனியலின் கிளைகள் (branches of morphology)

உருபனியலைச் சொல் திரிபு உருபனியல், சொல்லாக்கம் (சொல்வழி உருபனியல்) என்று இரு கிளையியல்களாகப் பகுக்கலாம். சொல் திரிபு உருபனியல் சொல்லன்களின் (lexemes) பலவடிவங்களைப் பற்றிக் கூறுகின்றது. ஆனால் சொல்லாக்கம் தந்துள்ள அடியிலிருந்து புதிய சொற்களை ஆக்குவது பற்றிக் கூறுகின்றது. சொல்லாக்கத்தை ஆக்கப்பாடு (derivation) என்றும்

கூட்டாக்கம் (compounding) என்றும் வகைப்படுத்தலாம். ஆக்கப்பாடு ஒட்டாக்கத்தால் புதிய சொல்லன்களை உருவாக்குவதைப் பற்றிக் கூறுகின்றது. கூட்டாக்கம் இரண்டு அல்லது இரண்டிற்கு மேற்பட்ட தகுதி வாய்ந்த பகுதிகள் சேர்ந்து புதிய சொல்லன்களை உருவாக்குவது பற்றிக் கூறுகின்றது. ஆக்கப்பாட்டை வகுப்பு மாறா ஆக்கப்பாடு, வகுப்பு மாறும் ஆக்கப்பாடு என இரு வகுப்புகளாகப் பிரிக்கலாம். அடியிலிருந்து அதே வகுப்பைச் சார்ந்த சொல்லன்களை ஆக்குவது வகுப்பு மாற்றாஆக்கப்பாடாகும் (எ.கா. *வேலை + ஆள் > வேலையாள்*); அடியிலிருந்து வேறுபட்ட சொல் வகுப்பைச் சார்ந்த சொல்லன்களை ஆக்குவது வகுப்பு மாற்றும் ஆக்கப்பாடாகும் (*அன்பு + ஆக> அன்பாக*, *வாழ்+கை> வாழ்க்கை*). கூட்டாக்கத்தையும் ஆக்கப்பட்ட கூட்டுகளின் வடிவ வகுப்பு அடிப்படையில் கூட்டுப் பெயர், கூட்டுவினை எனப் பிரிக்கலாம். ஆக்கப்பாடுகளைப் பொருண்மையியல் அடிப்படையில் உள்வட்டக்கூட்டு (எ.கா. மாம்பழம், மாமரம்), வெளிவட்டக்கூட்டு (எ.கா. *கால்நடை*), நேரியைபுக் கூட்டு, உம்மைத் தொகைக் கூட்டு எனப் பிரிக்கலாம் (எ.கா. செடி கொடிகள்). கீழே தந்துள்ள படம் உருபனியலின் அடிப்படையான பாகுபாடுகளைக் காட்டும்.



உருபு, உருபன், மாற்றுருபன் (Morph, morpheme, and allomorph

ஒலிகள் தனித்து நின்றோ பல சேர்ந்து நின்றோ சொற்பொருளையோ அல்லது இலக்கணப்பொருளையோ தருமாயின் அதனை உருபன் எனலாம். உருபன் என்பது மொழியில் காணப்படும் ஒரு பொருள் தரும் சிறிய அலகு ஆகும். சொல் என்பது ஒரு உருபனையோ அல்லது ஒன்றுக்கு மேற்பட்ட உருபன்களையோ கொண்டிருக்கும். உருபனியலின் அடிப்படைக்கூறு உருபன் ஆகும். ஒரு கூற்றை அல்லது சொல்லை பொருள் தரும் அலகுகளாகப் பிரிப்பது உருபனியல் ஆய்வாகும். இவ்வலகை மேலும் பிரிக்க இயலாது. இத்தகைய மேலும் பிரிக்க இயலாத பொருண்மை தரும் அலகுகளை உருபன்கள் என்று அழைக்கலாம். எவ்வாறு

ஒலியன்களுக்கு மாற்றொலிகள் உள்ளனவோ அது போலவே உருபனுக்கும் மாற்றுருபுகள் உள்ளன. எடுத்துக்காட்டாக பின்வரும் சொற்களை எடுத்துக்கொண்டால் *மரம்*, *மரங்-கள்*. *மரம்* என்பதும் *மரங்* என்பதும் மரம் என்ற உருபனின் மாற்றுருபுகள் எனலாம். ஒரு சொல்லை பொருண்மை தரும் கூறுகளாகப் பிரிக்கையில் அவற்றை உருபுகள் என்றுதான் அழைக்க வேண்டும். இவ்வுருபுகள் அவற்றின் செயல்பாடு மற்றும் வருகைமுறை அடிப்படையில் உருபன்களாகக் குழுமப்படும்.

உருபனியல் ஆய்வு (Morphological Analysis)

ஒரு மொழியின் உருபுகளையும் உருபன்களையும் உள்ளடக்கியப் பகுதி உருபனியல் எனப்படும். ஒரு உரையில் கையாளப்பட்டுள்ள மொழியின் உருபன்கள், உருபுகள், மாற்றுருபுகள் ஆகியவற்றை ஆய்வு செய்தல் உருபனியல் ஆய்வு எனப்படும்.

உருபனியல் மாற்றம் (Morphological Change)

காலப்போக்கில் ஒரு மொழியில் தோன்றும் மாற்றம் உருபனியல் மாற்றம் எனப்படும். எடுத்துக்காட்டாக பழந்தமிழில் வழக்கில் இருந்த தன்மை வினைமுற்று விகுதிகளான *-என்*, *-அல்*, *அன்* போன்றவைத் தற்காலத் தமிழில் இல்லாமல் இருப்பதைக் கூறலாம்.

உருபனியல் கட்டுப்பாடு (Morphological Condition)

உருபுகளால் விதிக்கப்படும் கட்டுப்பாடு உருபனியல் கட்டப்பாடு எனப்படும். எடுத்துக்கட்டாக தமிழ் மொழியில் *ஒரு* என்ற அடை மெய்யெழுத்தில் தொடங்கும் சொல்லுடனும் (ஒரு பெண்), *ஓர்* என்ற அடை உயிரெழுத்தில் தொடங்கும் சொல்லுடனும் (ஓர் அசை) வருதல் விகுதிகளில் *அர்* விகுதிக்கு முன்னர் வரும் *-அன்* - சாரியையை குறிக்கும். எடுத்துக்காட்டாக *செய்தனர்*.

உருபனியல் வகைப்பாட்டுக் கூறு (Morphological Cateory)

இலக்கண கூறுகள் கொண்ட உருபன்களை வகை வகையாகப் பிரித்தலைத் தான் வகைப்பாடு எனலாம். உருபனியல் வகைப்பாட்டுக் கூறு என்பது சொற்பொருள் அற்ற, இலக்கணப் பொருள் கொண்ட உருபன்கள் எனப்படும். எடுத்துக்காட்டாக தமிழ் மொழியில் பன்மை விகுதி *-கள்*. வேற்றுமை உருபுகள் *-ஐ-, கு* பேன்றவைகளைக் கூறலாம்.

உருபனியல் பாகுபாடு (Morphological Classification)

பாகுபாடு என்பது வேறுபாடு தெரியும் வகையில் பிரிக்கும் பிரிவு ஆகும். உருபனியல் பாகுபாடு என்பது சொற்களை அவை ஏற்று வரக்கூடிய விகுதிகளின் அடிப்படையில் வேறுபடுத்திக் காட்டிப் பாகுபடுத்துவது என்று பொருள்படும்.

அடிச்சொல் (stem)

அடிச்சொல் என்பதன் ஆங்கில நிகரன் stem ஆகும். அடிச்சொல் வேர்ச்சொல்லாகவும் (எ.கா. தமிழில் வினயடிகளான *படி*, *வா* என்பன) இருக்கலாம் வேர்ச்சொல்லும் வேறு உருபனும் சேர்ந்த ஒன்றாகவும் இருக்கலாம் (தமிழில் வினையடிகளான எ.கா. *படிப்பி*, வருவி என்பன).

உருபன்களின் வகைகள் (types of morpheme)

மரபு உருபனியில் உருபன்களை அவற்றின் இணையும் முறையின் அடிப்படையில் சேர்க்கை உருபன்கள் (additive morphme), குறைப்பு உருபன்கள் (substrative morpheme), பதிலி உருபன்கள் (replacive morpheme) என மூன்று வகைகளாகப் பிரிக்கும். இவை முறையே கூடுதல், குறைதல், பதிலாய் வருதல் என்ற மூன்று இணைப்புச் செயல்பாடுகளைக் குறிப்பிடும்.

சேர்க்கை உருபன் (Additive Morpheme)

மொழியில் வழங்குகின்ற ஒட்டுகள் அனைத்தும் சேர்க்கை உருபன்களே. அடிச்சொல்லுக்கு முன்னரோ பின்னரோ அல்லது அடிச்சொல்லிலோ இவை சேர்க்கப்படுகின்றன. எடுத்துக்காட்டாக, அ + நீதி > அநீதி, வாழ் + க்கை > வாழ்க்கை, im + possible > impossible, Arabic example: ktb + (i-a) > kitab (book) .

ஒன்றோடொன்று கூடும் நிலையை உடைய உருபன்பகள், வேர் உருபன்கள் (roots, முன்னொட்டுகள் (prefixes), பின்னொட்டுகள் (suffixes), உள்ளொட்டுகள் (infixes), அடுக்குகள் (reduplicatives) முதலியவை கூடுநிலை உருபன்களாகும். இவற்றில் நால்வகை ஒட்டுகளும் கட்டுண்ட உருபன்கள் (bound morphmes) ஆகும். கட்டுண்ட உருபன்கள் தனிச்சொல்லாக வரவியலாத உருபன்களாகும். தனிச்சொல்லாக வரவியலும் உருபன்கள் (எ.கா. தமிழில் *கை, கால்* என்பன) தனி உருபன்களாகும் (free morphemes),

முன்னொட்டுகள் எ.கா. அ-நீதி என்பதிள்ள அ-, un-happy என்பதிலுள்ள un-

பின்னொட்டுகள் எ.கா.: கால்-கள் என்பதிலுள்ள -கள், வாள்-ஐ என்பதிலுள்ள ஐ.

உள்ளொட்டுகள் எ.கா. கீழே தரப்பட்டுள்ளது.

முன்னொட்டு (Prefix)

அடிச் சொல்லுக்கு முன்னால் இணைக்கப்படும் ஒட்டுக்கு முன்னொட்டு என்று பெயர். faith-ful 'நம்பிக்கையான' un-faithful 'நம்பிக்கையற்ற' possible 'முடியும்: im-possible 'முடியாது' necessary 'தேவை': unnessary 'தேவையற்ற'; இவற்றில் un-, im- என்ற எதிர்மறை ஒட்டுக்கள் அடிச் சொல்லுக்கு முன்னால் சேர்க்கப்பட்டுள்ளதால் அவை முன்னொட்டுகள் எனப்படுகின்றன. தமிழில் முன் ஒட்டுக்கள் இல்லை. ஆனால் கடன் சொற்களில் முன்னொட்டுகள் சேர்ந்து இன்று

தமிழில் வழங்கப்படுகின்றன. எடுத்துக்காட்டாக நீதி > அநீதி, நாற்றம் > துர்நாற்றம், நியாயம் > அநியாயம், மலா > அமலா இவற்றில் 'அ' என்ற எதிர்மறை ஒட்டாகவும் துர்- என்பது 'கெட்ட' என்ற பொருளுடைய முன்னொட்டாகவும் அமைந்துள்ளன.

பின்னொட்டு (Suffix)

அடிச் சொல்லுக்குப் பின்னால் சேர்க்கப்படும் ஒட்டுக்குப் பின்னொட்டு என்று பெயர். இன்றைய தமிழில் பெரும்பான்மையான ஒட்டுகள் பின்னொட்டுகளாகவே அமைந்துள்ளன. வேற்றுமை உருபன்கள், பால் உருபன்கள், எச்ச உருபன்கள், கால உருபன்கள் எல்லாம் தமிழில் பின்னொட்டுகளாகவே அமைந்துள்ளன. எடுத்துக்காட்டு, படி+த்த்+ஆன்> படித்தான், வ+ந்த்+உ> வந்து, பார்+வை > பார்வை, நட+ ந்த்+ஆல் > நடந்தால்.

ஓரஒட்டுகள் (circumfixes)

முன்னொட்டும் பின்னொட்டும் இணைந்து ஒரு அடியுருபனை சூழ்ந்து ஒரு உருபனாகச் செயல்படும். இவ்வொட்டுகள் ஓரஒட்டுகள் (circumfixes) எனப்படும். ஓரஒட்டுகள் என்பதன் ஆங்கில நிகரன் circumfixes ஆகும். எடுத்துக்காட்டாக ஜெர்மன் மொழியில் உள்ள பின்வரும் ஓர ஒட்டுகளைக் கூறலாம்:

film-en 'to film' ge-film-t 'filmed

frag-en ' to ask ge-frag-t 'asked'

lob-en 'to praise' ge-lob-t 'parised'

zeig-en 'to snow' ge-zeig-t 'shown'

ge- என்பதும் - t என்பதும் இணைந்து இறந்தகாலத்தைக் குறிப்பிடும். இவை ஓர ஒட்டுக்களுக்கு எடுத்துக்காட்டாகும்.

உள்ளொட்டுகள் (Infixes)

உள்ளொட்டுகள் ஒரு தொடரா அடியை (தொடரா உருபனை= discontinous morpheme) உருவாக்கும். வியட்நாமில் பேசப்படும் சரவ் என்ற மொழியில் உள்ளொட்டுகள் காணப்படுகின்றன.

(Charau, a language of Vietnam)

sulat 'write'

s-um-ulat 'wrote

s-in-ulat 'was written'

-um-, *-in-* என்ற ஒட்டுகள் *sulat* என்ற உருபனை இரண்டாகப் பிரித்து உள்ளே செருகப்பட்டுள்ளன. இவை உள்ளொட்டுக்களுக்கு எடுத்துக்காட்டாகும்.

செருகு ஒட்டுகள் (interfixes)

தொடரா ஒட்டுகளையும் தொடரா அடிகளையும் உள்ளடக்கிய ஒரு தனித்தன்மையான உள்ளொட்டுகள் செருகு ஒட்டுகள் (interfixes) எனப்படும். இஜிப்சியன்-அரேபிய (Egyptian Arabic) மொழியில் இவ்வகையிலான ஒட்டுகள் காணப்படும். இதில் அடிகள் மெய்யெழுத்துகளின் சட்டகங்கள் போல இருக்கும். இதில் உயிரெழுத்துகள் ஒட்டுகளாகச் செருகப்படும்.

Egyptian Arabic

ktb ‘write’

‘katab ‘he wrote’

‘jiktib ‘he will write’

mak’taba ‘bookshop

ma’kaatib ‘book shops’

kaatib ‘clerk’

இடையொட்டு (Infix)

அடிச் சொல்லுக்கு உள்ளே அதாவது அதனுடைய வடிவத்திற்கு இடையில் சேர்க்கப்படும் ஒட்டுக்கு இடையொட்டு என்று பெயர். தமிழில் இடையொட்டு கிடையாது, அராபிக் மொழியில் இடையொட்டு உண்டு. எடுத்துக்காட்டு: ktb ‘to write’: katab ‘to write ‘kitab ‘book’. மெய்யெழுத்துக்களுக்கிடையில் செருகப்படும் உயிரெழுத்துக்கள் இடையொட்டுகள் எனப்படும்.

இரட்டையுருபன் (reduplicative morpheme)

அடியின் ஒரு பகுதி இரட்டுவதால் விளையும் உருபன் இரட்டையுருபன் (**reduplicative morpheme**) எனப்படும். முழு அடியும் இரட்டித்தால் அது கூட்டாக்கத்தை ஒக்கும்.

எ.கா. Afrikaans மொழியிலிருந்து

sumulat ‘to write su-sulat ‘will write’

bumasa ‘to read’ ba-basa ‘will read’

?umaral ‘ to teach ?a-?aral ‘will teach’

இணையொட்டு (Suprafix or Simulfix)

அழுத்தம் (stress, சுரம் (pitch) போன்றவற்றை உருபனாகக் கொள்ளும் போது அவை இணையொட்டு என்று வழங்கப்படும். *கண்ணன்* என்ற சொல்லை எந்தவித அழுத்தமும்

இல்லாமல் கூறும் போது இயல்பான பொருளையும் *கண்ணன்* என்று ஒரு வித அழுத்தத்தோடு கூறும் போது 'கண்ணனா' என்ற பொருளையும் தருகிறது. இந்த அழுத்தம் பொருள் வேறுபாட்டிற்குக் காரணமாக அமைவதால் இது இணையொட்டு என்று வழங்கப்படுகிறது.

ஆக்க ஒட்டு (Derivational Affix)

விவரண மொழியியலில் ஒரு விகுதி இணைந்து ஒரு சொல்லிலிருந்து மற்றொரு சொல்லை உருவாக்குதல் ஆக்கம் எனப்படும். எடுத்துக்காட்டாக *பாடு* என்ற அடிச் சொல்லுடன் *தல்* என்ற ஒட்டு இணைந்து *பாடுதல்* என்ற சொல் உருவாகிறது. இங்கு ஒரு சொல் வகையிலிருந்து மற்றொரு சொல் வகையை ஆக்கப் படுகின்றது; இவ்விதம் புதிய சொல்லை ஆகப் பயன்படும் ஒட்டு ஆக்க ஒட்டு எனப்படும். *செய்* என்ற வினையிலிருந்து *செய்கை* என்ற பெயரை ஆக்குவதற்கு *கை* என்ற ஒட்டு பயன்படுகிறது.

குறைப்பு உருபன் (Subtractive Morpheme)

சில மொழிகளில் அடிச்சொல்லில் உள்ள சில ஒலியன்கள் குறைக்கப்படுவதன் மூலமும் பொருள் வேறுபாடு கொண்ட புதிய உருபன்கள் உருவாக்கப்படலாம். குறைப்பு உருபன்கள் மொழியில் மிக அருகியே காணப்படுகின்றன. பொதுவாக வரலாற்று ரீதியில் பொருளை வேறுபடுத்துகின்ற ஒலி மாற்றம் ஏற்படுவதின் மூலம் இக்குறைப்பு உருபன்கள் நிலவுகின்றன. பின்வரும் எடுத்துக்காட்டில் ஆண்பால் உருபனை பெண்பால் உருபனிலிருந்து இறுதி ஒலியன் அலகு குறைக்கப்பட்டு ஆக்கப்படதாகும். எனவே ஆண்பால் உருபன்களை பெண்பால் உருபன்களின் குறைப்பு உருபன் எனலாம். பிரஞ்சு மொழியில் பெயரடை ஒவ்வொன்றிற்கும் ஆண்பால் பெண்பால் என இரு படிவங்கள் உண்டு. 'கெட்ட' என்னும் பொருளில் வரும் ஆண்பாலில் / movez / என்பதும் பெண்பாலில் /movɛ / என்பதும் வருகினறன. இவ்வாறே கீழ்வருவனவும் உள்ளன:

ஆண்பால்	பெண்பால்
/movɛz/	/move/ 'கெட்ட'
/cerΦz/	/cerΦ/ 'மகிழ்ச்சியான'
/šod/	/šo/ 'சூடான'
/gras/	/gra/ 'கொழுத்த'
/frɛš/	/frɛ/ 'புதிய'

பதிலி உருபன் (replacive morpheme)

பதிலி உருபன் என்பது அடிச்சொல்லின் ஒரு பகுதிக்குப் பதிலாய் வரும் உருபனாகும். எடுத்துக்காட்டாக ஆங்கிலத்தில் ஒருமை-பன்மை வடிவுகளான foot [fut]-feet [fi:t] என்பதிலுள்ள u -> i: பதிலி உருபனாகும்.

கட்டில்லா அல்லது தனி உருபன் (Free Morpgheme)

கட்டில்லா அல்லது தனி உருபன் என்பது அதற்கு முன்னரோ பின்னரோ வரும் உருபன்களைச் சார்ந்து வராமல் தனித்து வரும். அதாவது இதன் வருகை பிற உருபன்களால் கட்டுபடுத்தப்படாமல் தனிச் சொல்லாக வரும். இது பிற உருபன்களின் துனையின்றி தனித்து இயங்கி வருவதால் கட்டில்லா அல்லது தனி உருபன் எனப்படும்.. எடுத்துக்காட்டாக *நட, போ, வா, செல், மரம், செடி, மண், பூ* போன்ற சொற்கள் கட்டில்லா உருபன்களாகும். இவ்வுருபன்கள் தனித்து மொழியில் பயின்று வரும்.

கட்டுருபன் (Bound Morpheme)

கட்டுருபன் என்பது தனித்து இயங்காமல் அதற்கு முன்னரோ பின்னரோ வரும் உருபன்களைச் சார்ந்தே மொழியில் வரும். அதாவது இதன் வருகை பிற உருபன்களால் கட்டுபடுத்தப்படும். எடுத்துக்காட்டாக, இறந்காலத்தைக் குறிப்பிடும் *-த்த்-, -ந்த்-* (படி-த்த்-ஆன், வ-ந்த்-ஆன்) என்ற ஒட்டுக்களும் நிகழ்காலத்தைக் குறிப்பிடும் *-கிற்-, -கின்ற்-* (*படி-க்கிற்-ஆன், படி-க்கின்ற்-ஆன்*) என்ற ஒட்டுக்களும் எதிர்காலத்தைக் குறிப்பிடும் *-வ், -உம்* (*வரு-வ்-ஆன், வர்-உம்*) என்ற ஒட்டுக்களும், *அ* என்ற பெயரெச்ச ஒட்டும் (எ.கா. *ஓடிய* (*குதிரை*)) தனியாக வராது; ஒரு வினையுடன் சேர்ந்தே வரும். எனவே அவை கட்டுருபன்கள் எனப்படும்.

போர்ட்மென்டொ உருபன் (portmanteau morpheme)

ஒன்றுக்கு மேற்பட்ட பொருளுணர்த்தும் ஒரு கலவைத்தன்மையான உருபனை போர்ட்மென்டொ உருபன் (**portmanteau morpheme**) என்று கூறவர். எடுத்துக்காட்டாக இலத்தீன் மொழியில் ஒருமை, பன்மை பெருண்மைகளும் வேற்றுமை பொருண்மைகளும் கலந்த ஒரு கலவைத்தன்மையான உருபன்கள் பயன்படுத்தப்படுகின்றன. இவற்றை போர்ட்மெனடோ உருபன் என்பர். எடுத்துக்காட்டாக இலத்தீன் மொழியிலுள்ள ANNUS ‘year’ என்பதன் திரிபு பின்வருமாறு அமையும்:

	ஒருமை	**பன்மை**
எழுவாய் வேற்றுமை (Nominative)	ann-us	ann-i:
விளிவேற்றுமை (Vocative)	ann-e	ann-i:
செயப்படுபொருள் வேற்றுமை (Accusative)	ann-um	ann-o:s

உடைமை வேற்றுமை (Genitive)	ann-i:	ann-o:rum
கு-வேற்றுமை (Dative)	ann-o:	ann-i:s
நீங்கல் வேற்றுமை (Ablative)	ann-o:	ann-i:s

மேற்கண்ட எடுத்துக்காடுகளில் வரும் பின்னொட்டுகள் யாவும் ஒருமை அல்லது பன்மையையும் வேற்றுமையையும் ஒரே சமயத்தில் குறிப்பிட்டு நிற்பதை உணரலாம். இவ்வாறு ஒன்றுக்கும் மேற்பட்ட இலக்கணக் கூறுகளை அல்லது செயல்பாடுகளை வெளிப்படுத்தி நிற்கும் உருபன்களை போர்ட்மென்டொ உருபன் எனலாம். இதுபோன்று படித்தான் = படி-த்த்-ஆன், படித்தாள் = படி-த்த்-ஆள் என்ற சொற்களில் வரும் *ஆன்*, *ஆள்* என்பன பால் (ஆண்பால் பெண்பால்) வேறுபாட்டையும் ஒருமையையும் குறித்து நிற்கின்றன. இவ்வுருபன்களை போர்ட்மென்டொ உருபன்கள் எனலாம்.

கட்டாயமில்லா உருபன் (Optional Morpheme)

சில அமைப்புகளில் இத்தகைய உருபன்கள் வந்தமைந்தாலும் வந்தமையாவிட்டாலும் பொருள் மாறுபாடு எதுவும் தோன்றுவதில்லை. எனவே இதனை கட்டாயமில்லா உருபன் என்று கூறலாம். எடுத்துக்காட்டாக *நான் மரம் வெட்டினேன்*, *நான் மரத்தை வெட்டினேன்* என்ற வாக்கியங்களை எடுத்துக்கொண்டால், முதலாவது வாக்கியம் இரண்டாம் வேற்றுமை (-ஐ) இல்லாது வருகின்றது. இவ்வாறு கட்டாயம் வராமல் இருக்கும் உருபன் கட்டாயமில்லா உருபன் எனப்படும். இது போன்று *உடைய/அது* என்ற உடைமை வேற்றுமை உருபும் கட்டாயமின்றி வரல்லாம். எனவே அவையும் உருபன்கள் கட்டாயமில்லா உருபன் எனப்படும்: *அது எனது புத்தகம்*, *அது என்னுடைய புத்தகம், அது என் புத்தகம்.*

கட்டாய உருபன் (Obligatory Morpheme)

ஓர் உருபனியல் அமைப்பில் கட்டாயமாக அமைய வேண்டிய உருபன் கட்டாய உருபன் எனப்படும். எடுத்துக்காட்டாக, *நான் படித்தேன்* (*படி+த்த்+ஏன்* + *படித்தேன்*) என்ற வாக்கியத்தில் *ஏன்* என்ற உருபன் கட்டாயமாக வரவேண்டும். எனவே *ஏன்* என்பது கட்டாய உருபன் என அழைக்கப்படும். இது போன்று நாங்கள் செய்தோம் (செய்+த்+ஓம்= செய்தோம்) என்ற வாக்கியதில் *ஓம்* என்பது கட்டாய உருபனாகும்.

தொடரா உருபன் (Discontinuous Morpheme)

உருபனானது இடைவெளி விட்டு இரண்டு, மூன்று பிரிவாக அமையுமானால் அது தொடரா உருபன் என்று வழங்கப்படுகிறது. எடுத்துக்காட்டாக, *இராமனும் கண்ணனும்* என்ற

தொடரில் -*உம்*...-*உம்* என்பன தொடரா உருபனாகும். இது போன்று *சீதாவோ இராமனோ* என்ற தொடரில் -*ஓ*...-*ஓ* என்பன தொடரா உருபனாகும்.

உருபன் கட்டு மாற்றுருபு (morphologically conditioned allmorph)

ஒரு உருபனின் பொருள் வேறுபாடில்லாத உருபுகள் மாற்றுருபுகள் எனப்படும். இவை துணைநிலை வருகை முறையில் வரும். ஒரு மொழியில் ஒரு உருபன் வந்தால் மட்டுமே வரக்கூடிய ஒரு மாற்றுருபு உருபன் கட்டு மாற்றுருபு எனப்படும். எடுத்துக்காட்டாக, பழந் தமிழில் மொழியில் ஆண்பால் ஒருமை விகுதியான *அன், ஆன்* ஆகியவை மாற்றுருபுகள். இவற்றுள் *அன்* விகுதி *அன்* சாரியை வந்தால் மட்டுமே வரும். *வந்தனன்* (*வந் + அன் + அன்*); *ஆன்* பிற இடங்களீல் வரும் (வந் + ஆன்). *அன்* என்ற விகுதி *அன்* எனற சாரியையுடன் மட்டும் வருவதால் இது உருபன் கட்டு மாற்றுருபு எனப்படும்.

நிறைவு செய்யா உருபன் (Non-completive Morpheme)

உருபன் நிறைவு பெறாமல் வந்தால் அதாவது அது வரும் அமைப்பில் முழுமை பெறாமல் வந்தால் அது நிறைவு செய்யா உருபன் எனப்படும். எடுத்துக்காட்டாக, வ-ந்த் > வந்த்-, செய்-ஆ > செய்யா- என்பன நிறைவு செய்யா உருபன்களாகும்.

நிறைவு செய்யும் உருபன் completive Morpheme)

ஓர் உருபனியல் அமைப்பில் சில உருபன்கள் நிறைவு செய்யும் உருபன்களாக அமைகின்றன. அப்பொழுது தான் அவ்வமைப்பு முழுமை பெற்று விளங்கும். அதனையே நிறைவு செய்யும் உருபன்கள் என்று கூறுகின்றோம். எடுத்துக்காட்டாக, வ-ந்த்-ஆன் > வந்தான், செல்-ன்ற்-ஆர் > சென்றார், நட–ந்த்-அது > நடந்தது என்பனவற்றில் *ஆன், ஆர், அது* என்பன நிறைவு செய்யும் உருபன்களாகும். தமிழ் மரபிலக்கணம் இவற்றை வினைமுற்று விகுதி எனக் கூறும்.

இரட்டைக்கிளவி (Reduplicative Morpheme)

மரபிலக்கணத்தில் கிளவி என்பது சொல் என்பதைக் குறிப்பிடும். பிரித்தால் பொருள் தராததுமான ஒலிக் குறிப்பு போன்ற சொல் கிளவி எனப்படும். ஒரு குறிப்பிட்ட ஒலியன் வடிவம் அல்லது ஒலியன் சேர்க்கை இருமுறையோ அல்லது அதற்கு மேலோ வந்து ஒரு குறிப்பிட்ட பொருளை வெளிப்படுத்துமாயின் அது இரட்டைக்கிளவி எனப்படும். இவ்வாறு இரட்டித்து வரும் உருபனின் ஒரு பகுதியை அதிலிருந்து பிரித்தால் அவ்வடிவம் பொருளற்றதாகி விடுகிறது. எடுத்துக்காட்டாக, தமிழில் *சலசல* (சலசலத்தது), *பளபள* (*பளபளத்தது*), *படபட* (*படபடவென்று வெடித்தது*

4.3. கணினி சார் உருபனியல் ஆய்வு

கணினியைப் பயன்படுத்தி செய்யப்படும் உருபனியல் ஆய்வு கணினிசார் உருபனியல் ஆய்வு எனப்படும். கணினி மொழியியலின் முக்கியமான அடிப்படைச் செயல்பாடு எழுத்துக்களின் அல்லது ஒலியன்களின் கோர்வையை உள்ளீடாக எடுத்துக்கொண்டு ஒரு ஆய்வை/பகுப்பாய்வை வெளியீடாகத் தருவதாகும் (Trost, 2003: 25-47). உள்ளீடு செய்யும் கோர்வையை உள்ளுறையும் உருபன்களாக (undelying morphemes) அல்லது உருபனியல்-தொடரியல் பொருள்கோளாகக் (morphosyntactic interpretation) கண்டறிய (பொருத்த) இயலும்

எ.கா.

அ.Incompatiblities

ஆ. in+con+partible+ity+s

இ. incompatibility+NounPlural

அ-வை இ-ஆகக் கண்டறியும் எளிய வழி அகராதியில் முழு வடிவத்தையும் தருவதாகும். அதாவது இடது பக்கம் சொல்லையும் வலது பக்கம் அதன் பொருள்கோளையும் தரும் விதமாக இணைகளின் பட்டியலைத் தருவதாகும். இதன் நன்மை எளிமையும் எல்லாச் சாத்தியமான நிகழ்வுகளின் பிரயோகமும் ஆகும். இதன் பாதகம் மிகைகளும் (redundancy) அகராதியில் தரவியலாத வடிவுகளைக் கையாள இயலாத நிலையும் ஆகும்.

4.3.1. உருபனியலை மாதிரிப்படுத்தல்

கணிதம், இயற்பியல் மற்றும் கணினி அறிவியல் ஆகியவற்றைப் பயன்படுத்திச் சிக்கலான அமைப்புகளின் நடத்தையை உருவகப்படுத்தவும் படிக்கவும் கணினிகளைப் பயன்படுத்துவது கணக்கீட்டு/கணினி மாதிரியாக்கம்/மாடலிங் (computational modeling) ஆகும். ஒரு கணக்கீட்டு/கணினி மாதிரியில் ஏராளமான மாறிகள் உள்ளன, அவை ஆய்வு செய்யப்படும் அமைப்பை வகைப்படுத்துகின்றன (Wikipedia on computational model). உருபனியலைக் கணினிசார் மாதிரிப்படுத்துதல் (computational modeling of morpholgy) என்பது மொழியின் உருபனியல் அமைப்பையும் (morphological structure) உருபன்கள் வருமுறையையும் (morphotactis) உருபனியல் (morphology) மற்றும் உருபொலியனியல் (morphophonemics) விதிகளையும் கணியைப் பயன்படுத்தி மாதிரிப்படுத்துவதாகும்.

4.3.1.1. மொழியியல் கோட்பாடுகள்

மொழியியல் கோட்பாடுகள் உருபனியல் ஆய்வின் கட்டுமான அலகாக அல்லது அடிப்படை அலகாக உருபன் உருபனியல் கோட்பாடுகளின் பார்வையில் மிகவும் வேறுபடுகின்றன. உருபனியல் ஆய்வின் முக்கியமான செய்தி அடிப்படை உருபனியல் அலகின் தனிப்பட்ட மற்றும் ஒருங்கிணைக்கப்பட்ட செயன்மை நோக்குகளின் அடையாளம் காணலும்

உருப்படுத்தமும் ஆகும். ஹாக்கெட்டின் (Hockett, 1958) உருபனியலின் இரண்டு மாதிரிகளான சொல்லும் வரிசைமுறையும் (Item and Arrangement (IA)) மற்றும் சொல்லும் செயற்பாங்கும் (Item and Process (IP)) என்பன உருபனியலில் அமைப்பியல் நோக்கை உட்படுத்துகிறது. இவ்வர்த்தத்தில் ஒரு உருபன் ஒரு மொழியின் கூற்றின் மிகச்சிறிய தனிப்பட்ட பொருள் தருகிற தனிமம் ஆகும். சொல்லும் வரிசைமுறையும் மாதிரி இவை ஒன்று சேர்க்கப்படும் அணுகுமுறை (concatenative approach) ஆகும். இதில் உருபன்கள் சொல் அலகுகளாகும் (lexical units); உருபனியல் இம்மாதிரியான அலகுகளின் ஒட்டுநிலையாகும் (agglutination); சொற்கள் உருபன்களின் கிடைநிலை வரிசைத் தொடர்ச்சிகளாகக் கருதப்படுகின்றன. சில வடிவங்கள் பிற வடிவங்களுடன் பகுதி ஒலியியல் பொருண்மையியல் (partial phonetic-semantic resemblance) ஒற்றுமையைப் பிரதிபலிக்கும் என்பது இம்மாதிரியின் மையப் பார்வையாகும். ஒரு மொழியின் உருபொலியனியல் அமைப்பொழுங்கை விளக்குவது உருபன்களின் தெரிவடைவுகள் (items), இவ்வுருபன்கள் வரும் சாத்தியமான தொடர்ச்சிகள் (arrangement) மற்றும் மொழியின் ஒவ்வொரு உருபனும் மெய்ப்படுத்தப்படும் உருபுகளின் விளக்கம் (morphological link) என்பனவற்றை உள்ளடக்கும். சொல்லானது எச்செயற்பாங்கையும் உட்படுத்தாததாக உருபன்களின் அனுமதிக்கப்பட்ட தொடர்ச்சியாக ஆயப்படுகிறது. இவ்வாறு சொல்லும் வரிசைமுறையும் மாதிரி உருபன்சார்ந்த ஆனால் செயற்பாங்கு சாராத மாதிரியாகும்.

4.3.1.1.1. சொல்லும் செயற்பாங்கும் மாதிரி

இம்மாதிரி சொல் வடிவுகள் அமைப்பு மற்றும் செயல்பாடு (structure and function) அல்லது வடிவம் மற்றும் பொருண்மை (form and meaning) இவற்றின் இணைப் பண்புக்கூறுகளைக் கொண்டிருக்கிறது மற்றும் அத்தனிமங்கள் மாற்றங்கள் (transformations) வழி புதிய வடிவங்களை விளைவிக்க வேறுபட்ட மாறுதல்களை (modifications) அடைகிறது என்ற கருத்துருவாக்கத்தில் செயல்படுகின்றது. புதிய சொல் வடிவங்களை உருவாக்க வேண்டி உருபன்கள் மீதோ சொற்கள் மீதோ பயன்படுத்தப்படும் ஒரு குழும ஆக்கச் செயற்பாங்குகள் உருபனியல் என்று கருதப்படுகின்றது. எனவே சொல்லும் செயற்பாங்கும் மாதிரியில் ஒரே ஒரு அறியப்பட்ட உள்ளுறையும் அடிப்படை மாற்றுருபு (underlying basic allomorph) தான் இருக்கிறது. இது சூழல் அடிப்படையில் மாற்றுருபின் வடிவை எடுக்கும்.

4.3.1.1.2. சொல்லும் அடுக்கும் மாதிரி

சொல்லும் அடுக்கும் மாதிரி (word and paradigm model) மாறிகளின் அடுக்கிற்குள் (paradigm of variables) செயல்படும் சொல்லை அடிப்படையாக ஏற்கிறது. சொல்லும் அடுக்கும் மாதிரி ஒவ்வொரு சொல்லையும் முழுமையாகவும் அது வெளிப்படுத்தும் பண்புகளின் (characters) தனிப்பட்ட குழுமத்திற்கு அதன் ஒட்டுமொத்தமான விளைவையும் கவனக்குவிப்பு செய்கிறது என்ற அளவில் வேறுபட்டது. இம்மாதிரி கலவைத்தன்மையை விலக்குகிறது; ஏனென்றால் இது

ஒரு சொல்லை இரண்டு இணையான மேலுறல் செய்யாத அலகுகளாகப் பகுக்கிறது: ஒன்று வடிவத் தன்மங்களுக்காக (formal units) மற்றொன்று செயல்பாட்டுக் குணங்களுக்காக (functional properties). சொல் அடுக்குச் சட்டகத்தில் இது உருப ஒலியனியல் (morphophonemic) விதியாகும். ஒரு சொல்லுடன் தொடர்புடைய உருபனியல்-தொடரியல் (morphosyntactic) மற்றும் உருபனியல்-பொருண்மையியல் (morphosemantic) செயன்மைகளின் ஒன்றுசேர்ந்த குழுமம் ஆகும்; ஒலியனியல் (phonemics) விதியல்ல. பகுப்பாய்வு ஒலியனியல் நேர்வுகளுக்கு இணையாக உருபனியல் பண்புகளுக்கும் உருபனியல் வடிவங்களுக்கும் இடையிலான பொருத்தங்களின் இயல்பு அல்லது எண்ணிக்கையை எல்லைப்படுத்தும் முழுமையான கட்டுப்பாடுகள் இல்லை. ஸ்டம்ப் (Stump 1998) திரிபுக்குச் சொல் மற்றும் அடுக்கு அணுகுமுறையைப் பின்வருமாறு விளக்குகிறார்: X என்ற ஒரு பகுதிக்குப் (stem) பயன்படுத்த இயலும். ஒரு விதி X-உடன் தொடர்புடைய ஒரு குழும உருபனியல் - தொடரியல் பண்புகளாலோ X-இன் ஒலியனியல் வடிவாலோ ஒரு குறிப்பிட்ட உருபனியல் வகுப்பில் X-இன் உறுப்பினர் நிலையாலோ அல்லது இம்மாதிரியான காரணங்களின் ஏதாவது ஒன்றுகையாலோ கட்டுப்படுத்தப்படும். இங்கு திரிபின் விதிகள் குழுமங்களாகக் குறிப்பிடப்படுகிறது. ஒவ்வொரு குழுமமும் அதன் வகையும் ஒன்றிற்கு மட்டுமே பயன்படுத்தப்படுகிறது. ஆக்கமுறை இலக்கணம் இலக்கணக் கோட்பாட்டில் எடுத்துக்கூறத்தக்கச் சக்தி வாய்ந்த உபாயத்தை (device) உருவாக்கியது. அதாவது வேறுபட்ட மொழி மட்டங்களை இணைக்கும் கணிக்கும் கருவியான மாற்றங்களை (transformations) உருவாக்கியது. உருபனியல் செயற்பாங்குகள் மாற்றச் செயன்மைகள் (transformational operations) என்பதைவிட ஆக்க இயக்க நெறிகளாகக் (derivational mechanism) கருதப்பெற்றன. இது ஆக்கமுறை இலக்கணத்தில் உருபனியலின் தானியங்கு உட்கூறின் (autonomous component) தேவையை வெளிப்படுத்திய சொல்லியலார் கருதுகோளுக்கு (lexicalist hypothesis) முன்மொழிந்தது. இக்கோட்பாடு ஒரு மொழியின் இலக்கணத்தின் ஒழுங்குப்படுத்துதலைப் பற்றி இரண்டு மைய வாதங்களை முன்வைக்கிறது:

1. எல்லா உருபனியல் செயற்பாங்குகளும் சொல் உட்கூறால் (lexical component) ஆளுமை செய்யப்படும்
2. எல்லா உருபனியல் உண்மைகளும் சொல்லாக்க விதிகள் (word formation rules) என்ற சிறப்பான விதிகளால் கூறப்படும்.

ஹல்லேயினுடைய (Halle, 1973) ஆய்வுரை சொல்லியலாரின் கருதுகோளுடன் (Lexical Hypothesis) இயைந்த முதல்மாதிரியாகும். உருபனியலுக்கும் மொழி பேசுபவரின் உள்ளுணர்வுக்கும் இடையிலுள்ள தொடர்பைக் காட்ட ஹல்லே உருபன்களின் தெரிவடையையும், ஒரு குழு சொல்லாக்க விதிகளையும் (Word Formation Rules (WFR), சொல்லாக்க விதிகளால் ஆயப்பட்ட சாத்தியமான ஆனால் இல்லாத சொற்களைத் தள்ளி

வைக்கும் வடிகட்டியையும் (filter), எல்லா அடுக்கு வடிவங்களுடன் (paradigmatic form) சொற்சொருகலுக்குத் தயாராக இருக்கும் மொழியின் எல்லா ஆனால் உண்மையான சொற்களை மட்டும் கொண்ட அகராதியையும் கொண்டிருக்கும் மாதிரியை உருவாக்கினார். இதற்கு எதிராக இதனுடன் வேறுபட்ட சொல் அடிப்படை மாதிரி (word based model) என்று அழைக்கப்பட்ட மாதிரி, ஆரனோப் (Aronoff, 1985) என்பவரால் முன்மொழியப்பட்டது. ஆரனோப்பின் கருதுகோள் பின்வருவனவற்றை எடுத்துரைக்கின்றது: 1) எல்லா சீரான சொல்லாக்கச் செயற்பாங்கும் சொல் அடிப்படையிலானது; 2) புதிய சொல் ஏற்கனவே இருக்கும் தனிச்சொல்லுடன் சீரானவிதியை பயன்படுத்தி உருவாக்கப்படும்; 3)புதிய சொற்களும் ஏற்கனவே இருக்கும் சொற்களும் முக்கிய சொல் வகைப்பாடுகளின் (major lexical categories) அங்கத்தினர்களாக வரும்.

செல் வலைப் பின்னல் மாதிரி (word network model) (Sing and Ford, 1984) உருபனியலுக்கு ஒரு உறவுசார் அணுகுமுறையைப் (relational approach) பின்பற்றி அகராதியின் மீது உருபனியலின் சார்வை நிறுவுகின்றது. இவ்வணுகுமுறை உருபனியலை இரண்டு முதன்மையான கொள்கைகளால் விவரணை செய்கின்றது: 1) உருபனியல் பகுப்பாய்வின் அடிப்படை உருபன் அல்ல, சொல்லாகும் 2) சொற்கள் அலகுகளாக ஒவ்வொன்றுடனும் கலவைத் தன்மையான வலைப்பின்னலால் தொடர்பு கொண்டுள்ளன. சொல் வலைப்பின்னல் மாதிரி உருபனியலின் இயல்பு என்ன என்பதைப் பற்றிச் சில குறிப்பிட்ட அடிப்படையான அனுமானங்களைக் கொண்டிருக்கின்றது:

1) பன்முக உருபனியல்களின் (multiple morphologies) விலக்கல்; ஈரிணை வகைகயான திரிபுக்கும் ஆக்கத்திற்கும் (inflection vs derivation), ஒன்றுசேர்த்தலுக்கும் பிரித்தலுக்கும் (concatenative vs non-concatenative) அல்லது தொடர்ச்சியற்றதற்கும் (non-linear) இடையில் எந்தவித வேறுபாடும் செய்யப்படவில்லை.

2) உருபனியல் செயன்மை செய்யும் அடிப்படை அலகாகச் சொல்லைத் தீர்மானித்தல். வேர் (root), பகுதி (stem), சொல்லன் (lexeme) போன்றவைகளுக்கு எந்த உருபனியல் அந்தஸ்தும் தரப்படவில்லை.

3) ஒருங்கிணைந்த (unified) உருபனியல் செயன்மை மீது நிலைநிற்றல். உருபனியலுக்கும் ஒலியனியலுக்கும் இடையில் உருப்படுத்தத்தின் இடைப்பட்ட மட்டம் அனுமதிக்கப்படவில்லை. எந்த உருபனியல் செயன்மைகளுக்கும் ‘அடிப்படை’ (basic) அல்லது ‘ஆக்கப்பட்ட” (derived) வடிவங்களின் கருத்துருக்களைத் தருவது போன்ற சிறப்பான திசை இல்லை.

5) எந்த வகைப்பாடும் உருபனியலால் மட்டும் தீர்மானிக்கப்படவில்லை என்பது வலியுறுத்தப்படுகின்றது.

6) சொல்லின் உருபனியல் சீரியநிலையின் ஆதாரம் எந்த கட்டுப்பாடுகளோ வரையறைகளோ எடுத்துக்கொள்ளப்படவில்லை (எ.கா. பக்க இணைப்புக் கட்டுப்பாடு (adjacency condition) போன்றவற்றிற்கு இடமில்லை). உருபனியல் இங்கு வடிவ அடிப்படையிலும் பொருண்மையியல் அடிப்படையிலும் உறவுகொண்ட சொற்களின் ஆய்வாகக் கருதப்படுகின்றது. சொல் என்பதைக் கருத்தில் கொள்ள ஒரு வெளிப்பாடு மூன்று பண்புக் கூறுகள் கொண்டிருப்பதாய் பண்பாக்கம் செய்யப்படுகின்றது: 1) ஒலியனியல் வடிவு, 2) வகைப்பாடு மற்றும் 3) பொருண்மை. சொல் வலைப்பின்னல் மாதிரி மொழிகளைக் கடந்த பொதுமைகளைக் கைக்கொள்ளும் ஒரு தனியான ஒருங்கிணைந்த கோட்பாட்டை (single unified theory) முன்மொழிகிறது. இது உருபனியலில் உலகப் பொதுமைகளை முற்கோள் செய்கின்றது. உருபனியல் திட்டம் (Morphological strategy (MS)) அல்லது சொல்லாக்கத் திட்டம் (Word Formation Stategy (WFS)) சொற்களுக்கு இடையே உள்ள உருபனியல் உறவுகளை விவரணை செய்கின்றது.

4.3.1.2. உளமொழியியல் கோட்பாடு

4.3.1.2.1 மனச்சொற்களஞ்சியத்திலிருந்து உருபனியல் அமைப்பின் சேகரிப்பும் மீளப்பெறுதலும்

உருபனியலின் பொதுவான பார்வை ஒரு மொழியின் கலவைத் தன்மையான சொற்களின் அக அமைப்பை விளக்குவதாகும். இவ்வெல்லாச் சொற்களும் உண்மையிலேயே இல்லாதிருக்கலாம்; ஆனால் இவை உருபனியல் செயன்மைகளின் ஒரு குழும அமைப்பொழுங்கால் ஆளுமை செய்யப்படுகின்றன. சொற்களஞ்சியத்தைப் பற்றி பல பார்வைகள் கைக்கொள்ளப்படுகின்றன. ஒரு மொழியின் அகராதி குறித்த பொதுக் கருத்துக்கு மாறாக, அகராதி தனிப்பட்ட சுதந்திரமான உள்ளுறையும் மொழி உட்கூறாக நடத்தப்படுகிறது. இரு வழியிலும் சொற்களஞ்சியத்தை ஒரு மொழியில் இருக்கிற சொல் அடுக்குகளின் பட்டியல் என விவரணை செய்யலாம்; இவை பேசுபவரால் அறியப்பட்டிருக்க வேண்டும்; ஏனென்றால் அவை இடுகுறித்தன்மையான குறிகள்; சில வழியில் அனுமானிக்க இயலாதவை (Aronoff and Ashen, 1998).

காரண காரிய அடிப்படையில் உருபனியல் சாத்தியமான சொற்களிலும் சொற்களஞ்சியத்திலும் செயல்படுவதால் உண்மையில் உள்ளதாயிருக்கும் சொற்களில்தான் இரண்டையும் ஒன்றையொன்று இணைக்கும் பொதுவான பண்புகள் இல்லை என்று ஒருவர் அனுமானிக்கலாம். உண்மையிலேயே இரண்டு முக்கிய காரணங்களுக்காக இதைத் தவறான கருத்தாக நிரூபிக்க இயலும்:

1. உருபனியலும் அகராதியும் முதன்மையாகச் சொற்களைக் கையாள்கின்றன. மேலும் ஒரே மொழிச் செயற்பாடுகளைச் செய்கின்றன.
2. உருபனியலும் அகராதியும் அடிப்படையில் சுதந்திரமான உறவைப் பங்கிட்டுக் கொள்கின்றன.

ஆன்னோப் மற்றும் ஆன்சன் (Aronoff and Anshen 1968) என்பவர்கள் பின்வருமாறு கூறுகின்றனர்: எச்சொற்கள் இருக்கிறது மற்றும் எவை சாத்தியமானவை என்பதற்கு இடையிலான வேறுபாடுகள் தனிப்பட்டவரின் அகராதி மற்றும் உருபனியல் இவற்றினால் மட்டும் விவரணை செய்யப்படுகின்றது. உருபனியலும் அகராதியும் ஒன்றிலிருந்து ஒன்று தொடங்குகின்றது என்பது உருபனியல் அடிப்படையில் கலவைத் தன்மையான சொற்களின் உருவாக்கலால் மறைமுகமாகக் குறிப்பிடப்படுகின்றது. இது பெரும்பாலும் பேசுபவரின் மன அகராதியில் சேகரிக்கப்பட்டுள்ள உண்மையில் வரும் அடிப்படைச் சொற்களின் மீது உருபனியல் விதிகளின் பயன்பாட்டால் செய்யப்படுகின்றது. உருபனியல் விதிகளால் சொற்கள் ஆளுமை செய்யப்படும் மற்றொரு நிகழ்வு ஒழுங்கின்மையின் மரபுரிமையாகும் (inheritance of irregulaity). இதில் மரபுப் பொருண்மையியலை உடைய கலவைத் தன்மையான சொற்கள் அவற்றின் ஊகிக்கப்பட்ட அர்த்தத்திலிருந்து சிறிது மாறுபாட்டை வெளிப்படுத்துகிறது. உருபனியல் அகராதியைச் சார்ந்திருக்கிறது. இருப்பினும் உருபனியல் அடிப்படையில் கலவைத் தன்மையான அடிகள் இயல்பாகச் சொல் பதிவுகளாகும். உருபனியல் கோட்பாட்டில் சொல் புரிந்து கொள்கை செயற்பாங்குகளின் பல பார்வைகள் மைய அகராதியில் எவ்வாறு உருபனியல் அமைப்புகள் உருப்படுத்தம் செய்யப்பட்டுள்ளன என்ற கேள்விக்குச் சாத்தியமான தீர்வுகளாக முன்மொழியப்பட்டுள்ளன. இதைப் பற்றி பல கேள்விகள் எழுகின்றன.

மேக்குவீன் மற்றும் கட்லர் (McQueen and Cutler, 1998) என்போர் பின்வருமாறு கேட்கின்றனர்: மன அகராதி உருபனியல் உறவுகளைக் குறியாக்கம் செய்யுமாறு ஒழுங்கமைக்கப்பட்டுள்ளதா? எடுத்துக்காட்டாக, இரண்டு சொற்கள் சொற்களஞ்சியத்தில் குறியாக்கம் செய்யபட்டுள்ள ஒரே பகுதியை அல்லது ஒரே ஒட்டை பங்கிட்டுக் கொள்கின்றன என்பது உண்மையா? இது முதன்மையாக உருப்படுத்தம் பற்றிய கேள்வியாகும்: நீண்ட நாள் நினைவில் (long-term memory) சொற்களின் அக அமைப்பு பற்றிய எத்தகவல் சேகரிக்கபட்டுள்ளது? எவ்வாறு? பேச்சு அல்லது எழுத்து உள்ளீட்டிலிருந்து மன அகராதிக்குப் புலனறித் தகவலின் பொருத்தவரைவின் (mapping) செயற்பாங்கில் உருபனியல் அமைப்பு எப்பங்களிப்பைச் செய்கின்றது? மேலும் ஆய்வு மற்றும் உருப்படுத்தம் இவற்றைப் பற்றிய கேள்விகளும் இருக்கின்றன. எவ்வகையிலான உருபனியல் பகுப்பாய்வு நடக்க வேண்டும்; மற்றும் அணுகுதல் உருப்படுத்தங்களின் (access representation) எவ்வடிவம் இம்மாதிரியான செயற்பாங்கின் விளைவாக இருக்க வேண்டும்?

சொற்களைப் பெறுகையில் மனிதன் சொற்களைப் பகுத்தாய்கின்றானா? அப்படியென்றால் எக்கோட்பாடு அடிப்படையிலான, தர்க்க அடிப்படையிலான அல்லது புள்ளியியல் அடிப்படையிலான அனுமானங்கள் இவ்வாய்வுகளை உள்ளுறை செய்யும்? இவை மற்றொரு குழுமக் கேள்விகளாக அமைகின்றன. எவை இவற்றிற்குத் தீர்மானமாக விடையளிக்கும் அல்லது கேள்விப் பற்றிய தீர்மானமாக எதை நிறுவ இயலும், எத்தனை பேசுபவர்

இருக்கிறார்களோ அத்தனை உருபனியல்கள் இருக்கின்றனவா என்பனவற்றைக் கவனத்தில் கொள்ளாமல், விதி அமைப்பொழுங்குகள் விதிவிலக்கு இல்லாதவையாக இருப்பதன் காரணமாக, உருபனியல் பல எண்ணிக்கையிலான சொற்களஞ்சியங்களைக் கடந்த ஒரு பொதுவான உலகப்பொது சட்டகத்தைத் தருகின்றது. ஒரு சொற்களஞ்சியத்தைத் தனிப்படுத்தப்பட்ட கருத்துருவாக்கமாகக் (individualized conception) கருத இயலும் என்றாலும், உருபனியல் செயற்பாங்குகள் வேறுபட்ட சொற்களஞ்சியங்களைக் கடந்து வெளிப்படும் பொதுமையாக்கங்களைக் கண்டுகொள்ளும். சொற்களஞ்சியத்தில் காணப்படும் வேறுபாடுகள் உருபனியலில் காணப்படும் வேறுபாடுகளாகக் கருதப்படத் தேவையில்லை.

சொல்லை அணுகுதலில் உருபனியலின் பங்களிப்பைப் பற்றிய வேறுபட்ட அணுகுமுறைகளை ஆய பின்வரும் உருபனியலின் உளவியல் மாதிரிகளைக் கருத்தில் கொள்ளலாம்.

4.3.1.2.2. மாதிரிகளின் வகைகள்

உருபனியலின் உள மொழியியல் மாதிரிகள் தொடர்பாகப் பல கேள்விகளை எழுப்ப இயலும். பெறுதலில் (access) உருபனியல் அமைப்பின் புலனறிவு ஏதாவது பங்களிப்பு செய்கிறதா அல்லது அதிலிருந்து வருகிறதா? பகுதிகளும் (stem) ஒட்டுகளும் (affixes) ஒரே அகராதியில் ஒன்றுடன் ஒன்று இணைந்து சேகரிக்கப்பட்டு மீளப்பெறப்படுகிறதா அல்லது வேறுபட்ட பதிவேடுகளும் (records) மற்றும் செயற்பாங்குகளும் உட்படுத்தப்படுகிறதா? அகராதியின் இயல்பு மற்றும் அதில் சொல் வடிவுகளின் உருப்படுத்தம் இவற்றைப் பொருத்தவரையில் சொற்களின் எழுதப்பட்ட பேசப்பட்ட, படிக்கப்பட்ட மற்றும் கேட்கப்பட்ட வடிவங்களுக்கு இடையே உள்ள வேறுபாடுகளை இல்லாமல் செய்யும் படிக்குச் சேகரிப்பு அருவத் தன்மையானதா? அல்லது அதைத் தனித்தனியான தொகுப்புகளாக ஆனால் ஒன்றுக்கொன்று இணைக்கப்பட்ட வல்லுனரின் சொற்களஞ்சியங்களாகக் கருதலாமா?

இவைகள் எல்லாம் சொல் பெறுகையில் உருபனியல் அமைப்பின் அந்தஸ்துடன் (படிநிலையுடன்) தொடர்புபடுத்தும் முக்கியமான கேள்விகளாகும். சொற்களஞ்சியத்திலிருந்து எவ்வாறு ஒரு மொழியின் சொற்களெல்லாம் பெறப்படுகிறது என்ற கேள்வி எவ்வாறு மனித மனம் கலவைத் தன்மையான சொற்களைப் பகுக்கிறது என்பதைத் தீர்மானிக்கும் ஆற்றலுள்ளதாகும். மன உருபனியலைப் பரிசோதிக்கும் முயற்சியில் சொல் புரிந்து கொள்கையின் ஒருசில முக்கியமான உளமொழியியல் மாதிரிகள் ஆராயப்பட்டன. சொல் பெறுதலின் வேறுபட்ட கோட்பாடுகள் முன்மொழியப்பட்டன. கீழே இவற்றைப் பற்றி விளக்கப்படுகிறது.

கோர்வை வருடும் மாதிரிகள் (String-Scan Models)

மாற்ற வலைப்பின்னல் மாதிரி (Transition Network Model)

ஃபாஸ்டர் (Forstor, 1976) தேவைப்படுகிற கோர்வைக்கு மனித முயற்சி சார் அகராதித் தேடலின் கொள்கை அடிப்படையில் அமைந்த மன அகராதியிலிலிருந்து ஒரு சொல்லை மீட்டெடுப்பதற்கு முயலும் சொல் அணுகலின் நேரடி மாதிரியைப் பரிந்துரைக்கிறார். இவ்வியக்க

நெறி மாற்ற வலைப்பின்னலாகும் (Forstor, 1976). இது இடது பக்கத்திலிருந்து வலது பக்கத்திற்கு கோர்வையை வருடுகிறது (ஆய்கிறது). கோர்வை தீர்வது வரை ஒரு நேரத்திற்கு ஒரு எழுத்து என்ற விகிதத்தில் தொடர்ச்சியாக முன்னேறும். ஒவ்வொரு நிலையிலும் பொருத்தம் காணப்பட்டால் மாற்றம் (transition) முன்னேறும். தேவைப்படும் நேரங்களில் பின்செல்லுகை (back tracking) நிகழும். கோர்வைக்கான தேடல் அகராதித் தேடலின் (dictionary search) அடிப்படையில் அமைந்தது; இதில் சாத்தியமில்லாத வெளியீடு ஒலியனியல் கட்டுப்பாட்டால் நீக்கப்படும். முன்னேற்றம் சரியான வழியில் இருந்தால் சொல்லிற்கான முற்றுநிலை வலைப்பின்னல் தீர்மானிக்கும். சிக்கனம் பின்வருவனவற்றில் காணப்படுகிறது.

1. வழிகளின் ஒன்றிணைப்பு (Merging of pathways): வழிகளின் ஒன்றிணைப்பு சொல்லில் ஒவ்வொரு இடத்திற்கும் ஒரு எழுத்தின் ஒரு தனி நேர்வைக் கூறுவதால் பெறப்படும் எழுத்துத் தனிமங்களின் அமைப்பு இருப்பிடம், குறிப்பிட்ட பதிவுகளின் இருப்பை விவரணை செய்யத் தேடல் வழிகளை அனுமதிக்கும்.

2.வழிகளின் எண்ணிக்கை (Number of pathways): வலைப்பின்னல் வழியிலான குறிப்பிடப்பட்ட வழிகளின் நேர்வுகள் எல்லாச் சாத்தியமான மாற்றங்களைவிடக் குறைவாகும்.

3. வழிகளின் அமைப்பு (structure of pathways) வழிகளின் அமைப்பு வருடல் முடிவுக்கு வருவதற்கு முன்னரே சொல் வடிவைச் சரியாகக் கணிக்கும்.

பின்வருவன ஃபாஸ்டர் கண்டறியும் இம்மாதிரி நேரிடும் அனுகூலமின்மைகள் ஆகும்:

1. இல்லாத சொற்கள் (non-words) மிக எளிதில் கண்டுகொள்ளக் கூடியவை.

2. நீண்ட சொற்களை அடையக் கூடுதல் காலம் எடுக்கும்.

இம்மாதிரி சொல் நிகழ்வெண் காரணியை கணக்கிலெடுக்கும் சொல் அணுகலின் ஒரு நோக்கை விளக்கத் தவறிவிட்டதாகக் கண்டறியப்பட்டுள்ளது. இம்மாதிரி சொல்லைக் கண்டுபிடிக்கும் மாதிரியாக இல்லாமல் ஒலியனை கண்டுபிடிக்கும் மாதிரியாகத் தொpகிறது. மேலும் கார்மென் (Garman, 1990:262) என்பவர் இவ்வியக்க நெறியின் குறைபாடாகக் குறிப்பிடும் அவர் பார்வையைக் கருத்தில் கொள்ளாமல் விட்டுவிடலாகாது: அகராதி தனியாகப் பட்டியலிடப்பட்டுள்ள எல்லா அலகுகளும் முழுவதுமாக உச்சரிக்கப்பட்டு தரப்படும் ஒரு புத்தக வடிவு ஆகும்; இது பக்கங்களைப் புரட்டுவதும் அவற்றின் பொருளடக்கத்தை வருடுவதும் எனப்படும் பயன்பாட்டின் எளிமை அடிப்படையில் கூடுதல் மிகையைக் கொண்டிருக்கிறது; மனச் சொற்களஞ்சியத்தைப் பொறுத்தவரை இம்மாதிரியான ஒழுங்குமுறை தேவையில்லை மற்றும் கூடுதல் விலைமதிப்புள்ளது.

4.3.1.2.2.1 தனித்தியங்கும் தேடல் மாதிரி (Autonomous Search Model)

நேரடி மாதிரியின் தோல்விக்குப்பின்னர் ஃபாஸ்டரின் (1979) மற்றொரு மாறுபட்ட மாதிரி இருநிலை ஆய்வு மாதிரியாகும் (two stage processing model); இது செயன்மையின்

தொடர்ச்சியான பாங்கைக் கொண்டது; இதில் முதல் நிலை இரண்டாவது நிலை தொடங்குவதற்கு முன் முடிந்துவிடும். முதல் நிலையில் தேடல் சொல் அலகுகளின் நிகழ்வெண்ணால் நிரல்படுத்தப்பட்டுள்ளது. முன்னர் கூறிய மாதிரி போல் இல்லாமல் இது பௌதிக அமைப்பைவிட நிகழ்வெண் தரநிரலில் (frequency rank ordering) கவனக்குவிப்பு செய்கிறது. உள்ளீட்டுக் கோர்வையின் வருடலைவிட செயன்மையின் இரண்டாவது நிலையில் அருவ இடக்குறி (abstract location marker) முதல் நிலையில் பெறப்பட்ட குறிப்பு அடிப்படையில் சொல் அலகின் அணுகலைப் பெற உதவுகிறது. இந்நெறிமுறை பட்டியல் வரிசை (cataloguing) போன்றதாகும். ஒரு முக்கிய கோப்பு (master file) வரிசைப்படுத்தப்பட்ட தகவலைக் கொண்டிருக்கிறது; இது கூடுதலான அணுகலுக்கும் பெறுகைக்கும் உதவுகிறது. தனித்தியங்கும் தேடல் மாதிரி பின்வரும் குறிப்புகளைக் கொண்டிருக்கிறது:

1. கூடுதல் நிகழ்வெண் உள்ள சொற்கள் குறைந்த நிகழ்வெண் உள்ள சொற்களைக் காட்டிலும் விரைவாக அணுகப்படும்.
2. இல்லாத சொற்களை விட்டுவிடுதல் உண்மையான சொற்களை ஏற்றுக்கொள்வதைக் காட்டிலும் கூடுதல் காலம் எடுக்கும்.

இருப்பினும் இருமட்ட மாதிரியின் கோட்பாடு மற்றும் தர்க்கக் கொள்கைகள் மேம்பட்டவைகள் அல்ல என்று பரிசோதனைகள் காட்டுகின்றன.

4.3.1.2.2.2 லோகோகன் மாதிரி (Logogen Model)

மார்ட்டனின் (Morton, 1969) ஆய்வி ஒலியனியல் ஆய்வு மற்றும் சொல் அணுகல் இவற்றிற்கிடையே உள்ள தொடர்பின் உதவியுடன் ஒவ்வொரு நிலையிலும் தோன்றும் உள்ளீட்டுக் கோர்வையின் சொல் அடையாளம் காணுகையைப் பற்றிய முன்னேற்றமான அனுமானங்களைச் செய்கிறது. தேர்வு அடிப்படையில் ஒத்திசையுள்ள பதில் குணங்களின் (selectively tuned response characteristics) அடிப்படையில் சொற்கள் மெய்ப்படுத்தம் செய்யப்படுகின்றன. இவ்வொழுங்குமுறை குறிப்பிட்ட மொழியியல் பண்புகளைச் சார்ந்திருப்பதுடன் மொழிக்கு வெளியே உள்ள குறியீடுகளிலிருந்தும் சொற்களின் இருப்புக்கு குறியீடுகளை எடுக்க இயலும். இவைகள்தான் லோகோகென்ஸ் (Logogens) ஆகும்; அதாவது புலனுணர்வு மற்றும் பொருண்மை உள்ளீட்டிற்குப் பதிலளிக்கும் ஒத்திசைக்கப்பட்ட புலனுணர்வு சார் உபாயங்கள் ஆகும். புலனறிவு மற்றும் சூழல் உள்ளீடுகள் ஊடாட்டம் செய்து புலனறி ஒழுங்குமுறைக்கு (cogniative system) வெளியீடுகளைத் தருகின்றன.

லோகொகென் மாதிரி கூடுதல் நிகழ்வெண் உள்ள சொற்களுக்குக் குறைந்த எல்லை இருக்கிறது, எனவே அணுகலைப் பெறுவதற்கு குறைந்த செயற்பாங்கு போதுமானது என்று குறித்துக்காட்டுகின்றது. ஃபாஸ்டர் இம்மாதிரியில் வடிகட்டல் நுட்பத்தின் திறனை வினவுகிறார். அதாவது உயர்ந்த நிகழ்வெண் உள்ள அலகுகள் சரியான மற்றும் தவறான பதில்களாகக் கூடுதலாகக் கிடைக்கக் கூடியதாய் இருக்கின்றன. சொற்களின் செவிப்புல மற்றும் கட்புல

பண்புகளின் மீது குவியம் செய்யும் உண்மையான சொல் கண்டுபிடிப்பு மாதிரியாக (word-detection model) இது காணப்பட்டிருந்தாலும் அவற்றின் சூழல்கள் மற்றும் சொற்களின் நிகழ்வெண் தரம், தவறான ஆய்வுகளை விட்டுவிடும் திறமின்மை என்பன இம்மாதிரியைக் கூடுதல் மதிப்புள்ளதாகச் செய்யவில்லை.

4.3.1.2.2.3 கோகோர்ட் மாதிரி (Cohort Model)

இது மார்ஸலன்-வில்சென் மற்றும் வெல்ஷ் (Morslen-Wilson & Welsh, 1978) என்பவர்களால் உருவாக்கப்பட்டது. இது மாற்ற வலைப்பின்னல் (வசயஇளவைடிஇ இநவறடிசம) மற்றும் லோகோகென் மாதிரிகளின் மீதான முன்னேற்றமாகும். இதில் ஊக்கப்படுத்தும் மட்டங்களை மிகச்சரியாகக் கூற இயலும் அறியப்படாத விகிதங்களில் லோகோகென்சின் பகுதி (முழுமையற்ற) ஊக்கத்திற்குப் பதிலாக ஊக்கப்படுத்தலின் இரண்டு அளவுகளின் இருப்பு மட்டும்தான் கூறப்பட்டுள்ளது (பூஜியம் அல்லது முழுமை) கோகோர்டின் ; (Cohort) எல்லா உறுப்பினர்களின் முழு விளக்கமும் உள்ளீட்டுத் தொடர்ச்சியின் முதல் அலகின் அடிப்படையில் நிகழ்கிறது; அதன் பின்னர் தனியான உயர் எல்லை மதிப்புகளுக்குப் பகுதி ஊக்கங்களின் தொகுப்புக்குப் பதிலாகத் தொடர்ச்சியின் இடமிருந்து வலமான பகுப்பாய்வு தொடரும்போது முழு ஊக்கப்படுத்தப்பட்ட அலகுகளின் முன்னேற்றமான நீக்கம் (progressive elimination இருக்கும். புரிதலின் இடம் கோர்வையின் இறுதியால் அல்லாமல் கோகோர்ட்டிலிருந்து இறுதி மாறுபட்ட எழுத்தின் நீக்கலால் உறுதி செய்யப்படும் முந்தைய மாதிரிகளின் கலப்புப் பண்புக்கூறுகள் (hybrid features) முக்கிய தொடர்புபடுத்தல்களை விளைவிக்கும்.

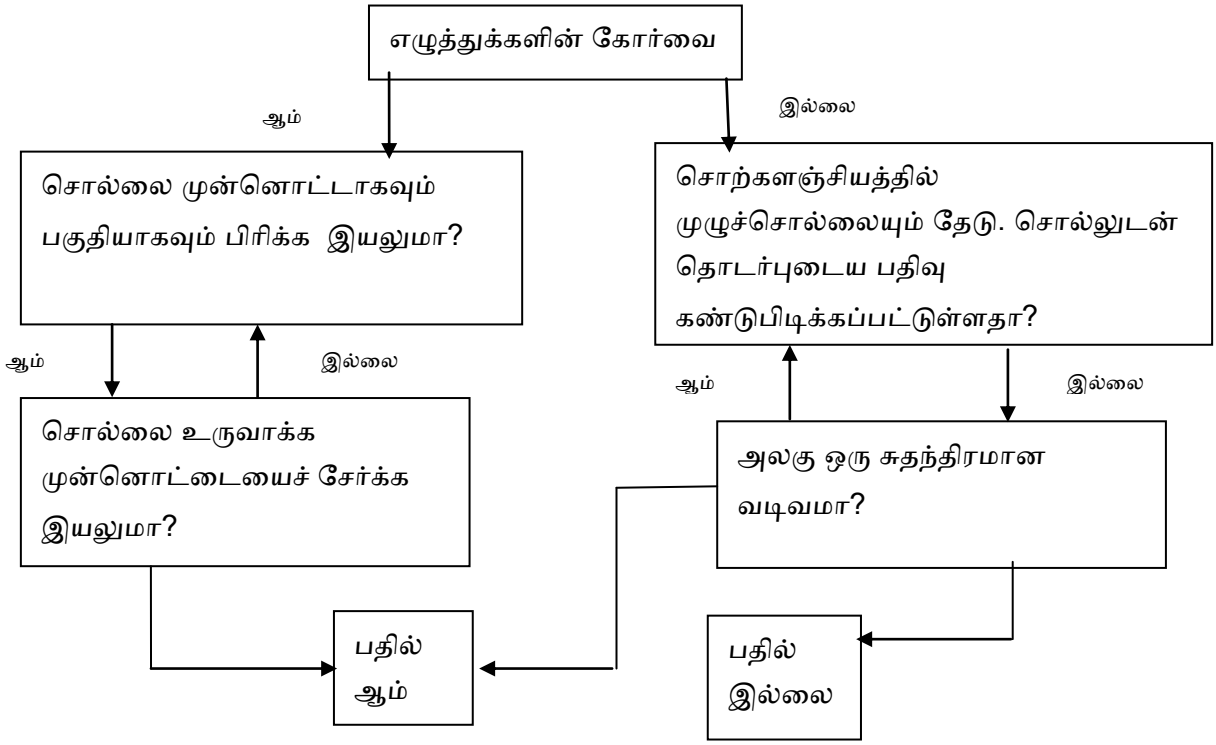
1. இது செவிப்புலக் குறியீட்டின் தொடர்ச்சியான இயல்பைச் சொல் புரிந்து கொள்ளலின் இயல்புக்குள் நேரடியாக உள்ளடக்குகின்றது.
2. மொழியின் எல்லாப் பிற சொற்களிலிருந்தும் (மற்றும் கோகோர்ட்) தொடர்ச்சியான அமைப்புகளில் அவ்வலகு விலகும் தனிப்பட்ட இடம் இருக்கிறது.

கோகோர்ட் மாதிரி குனுத்தில் (Kunuth, 1973) விளக்கப்பட்டுள்ளது போல் எழுத்து-கிளையமைப்புக் கோட்பாட்டை (letter-tree theory) பின்பற்றுகிறது. இருப்பினும் கோட்பாடு அடிப்படையில் இம்மாதிரியின் இயக்க நுட்பமும் அனுமானங்களும் கேட்ஸ் (Katz, et al, 1987) போன்ற கோட்பாட்டாளர்களால் தீர்மானிக்கப்பட இயலவில்லை.

4.3.1.2.2.4. வகைப்பாடு-பிளவு மாதிரிகள் (Category-Split Models)

டாஃப்ட் மற்றும் ஃபாஸ்டர் மாதிரியின் ஹண்டர்சனின் திறனாய்வு முதல்முறையாக ஒரு உளமொழியியல் மாதிரி சொல் புரிதலில் உருபனில் பகுப்பாய்வின் பங்கை வெளிப்படுத்தியது; குறிப்பாக டாஃப்ட் மற்றும் ஃபாஸ்டர் (Taft and Forster, 1975) என்பவர்களால் உருவாக்கப்பட்ட எழுதப்பட்ட சொல்லின் புரிந்துகொள்ளலுக்கான மாதிரி இங்கு கருத்தில் கொள்ளப்பட வேண்டும்.

இப்பகுப்பாய்வு சொல் அணுகலுக்கு முன்னரே பகுதிகளிலிருந்து ஒட்டுகளை நீக்குவதை உட்படுத்தும். இதன் செயல்பாட்டில் முக்கியமானது சொற்களின் எல்லைகளைக் கண்டுகொள்வது ஆகும். பின்வரும் ஒழுக்கு வரைபடம் (Flow chart) இம்மாதிரியின் செயல்பாட்டை எடுத்துக் காட்டும்.



எழுத்துக்களின் கோர்வை ஆம் இல்லை ஆம் இல்லை ஆம் இல்லை இல்லை சொல் புரிந்து கொள்கையில் உருபனியல் பகுப்பாய்வு (டாஃப்ட் மற்றும் ஃபாஸ்டர் (1975:644) இலிருந்து கார்னால் எடுத்தாளப்பட்டது. இம்மாதிரி தொடர்ச்சியான ஒன்று; இது நிரல்படுத்தப்பட்ட முறையில் ஒரு குழும மட்டங்களின் மீது செயல்படுகிறது. ஒரு பகுதி (stem) அதன் ஒரு தனி நிகழ்வால் உருப்படுத்தம் செய்யப்பட வேண்டியிருப்பதால் இவ்விதமான கூறிடல் மாதிரி சேகரிக்கப்பட்ட உருப்படுத்தங்களின் சிக்கனத்தை அனுமதிக்கிறது. ஒவ்வொரு மட்டத்திலும் நீக்கம் செயல்படுகிறது; மீதிப் பகுதியின் அடிப்படையில் சொல் தேடல் தொடர்கிறது. இம்மாதிரி ஒரு குழும பகுதி – உருபன் பதிவுகள் (stem-morpheme entries) மற்றும் நீக்கப்பட்ட ஒட்டுகளால் செயலூக்கம் செய்யப்பட்ட ஒரு குழும இணைப்பு விதிகள் (composition rules) இவற்றை வேண்டுகிறது. இது மறுபடியும் குறிப்பிட்ட சொல் ஆக்கங்களின் பண்புகளைத் தீர்மானிக்கிறது.

4.3.1.2.2.5 பெரிதாக்கப்பட்ட வேண்டப்பட்ட உருபனியல் (augmented addressed morphology)

இம்மாதிரி உருபனியல் கூறிடல் மூலம் சொல் அணுகலைச் செய்வதை நம்பும் டாஃப்ட் மற்றும் ஃபாஸ்டர் என்போரின் மாதிரிக்கு எதிரிடையாது. காரமாஸா (Caramazza, et al, 1998) என்போர் பெரிதாக்கப்பட்ட வேண்டப்பட்ட உருபனியல் கோட்பாட்டை ஆதரிக்கின்றனர். இது உருபனியல் அடிப்படையில் கலவைத் தன்மையுடைய சொற்கள் மன அகராதியில் ஒரே நேரத்தில் முழுச் சொல்லையும் அதன் உறுப்பு உருபன்களையும் தூண்டும் என்ற பார்வையை ஆதரிக்கும். சொல் அணுகல் முதலில் எதிர்கொள்ளப்படும் எந்த வடிவாலும் பெறப்படும். இது உருபனியல் கூறிடல் வழி சொல் அணுகலை நம்பும் டாஃப்ட் மற்றும் ஃபாஸ்டர் மாதிரிக்கு எதிரிடையாக இருக்கிறது. பரிசோதனைகள், செய்யப்பட்ட அனுமானங்கள் விளைவுடன் ஒத்திருக்கவில்லை என்று காட்டுகின்றன.

4.3.1.2.2.6. துணைக்கோள் பதிவுகள் மாதிரி (Satellite Entries Model)

லுக்காட்டேலா போன்றோர் (Lukatela et al, 1988) துணைக்கோள் பதிவுகள் மாதிரியைக் கைக்கொள்கின்றனர். தொடர்பு முறையில் (relational manner) எல்லாப் பிற மாற்றுருபுகளையும் ஒன்றுசேர்க்கும் ஒரு அடிப்படை மாற்றுருபை இதன் கருதுகோள் கூறுகிறது. இது உருபனியலின் சொல்லடுக்கு கோட்பாட்டிலிருந்து சிறிது வேறுபடுகிறது. அதாவது அடிவடிவிலிருந்து மாற்றுருபன்களை ஆக்குவதற்குப் பதிலாகச் சொல் வலை அமைப்பு போல ஒவ்வொன்றும் உறவு அடிப்படையில் இணைக்கப்படுகிறது.

ஸ்ப்ரோட் (Sproat 1992) என்பவர் லுகாட்டெ மாதிரியை உருபனியல் பகுப்பாய்வு மாதிரிக்குள் அடக்கலாம் என்ற கருத்தைக் கொண்டிருக்கிறார். அடிப்படை மாற்றுருபுக்கு முக்கியத்துவம் தரப்பட்டு அதிலிருந்து எல்லாப் பிற மாற்றுருபுகளும் உறவுப்படுத்தப்படுகின்றன.

4.3.1.2.2.7 தனி-பதிவுகள் மாதிரி (Separate- Entries Model)

மற்றொரு தனி-பதிவுகள் மாதிரி (Andrews 1986; Fowler 1985; Grainger et al 1991; Schreuders et al 1990; Schriefers et al 1991, 1992)) ஒரு சொல்லின் வேறுபட்ட வடிவுகள் அவற்றின் புற வடிவம் வழி பெற்ப்பட வேண்டும் என்று நம்புகிறது. இது வேறுபட்ட புற வடிவுகளின் நிகழ்வெண்ணை அடிப்படையாகக் கொண்டது.

4.3.1.2.2.8 எளிய கூறிடப்பட்ட மாதிரி (Naïve Decompositional Model)

எளிய கூறிடப்பட்ட மாதிரி பகுதிகள் மற்றும் ஒட்டுகள் வழி சொல் அணுகலைச்செய்யும் டாஃப்ட் மாதிரி வழியில் அமைந்தது.

4.3.1.3 உளமொழியியல் கண்டுபிடித்தங்களின் குறிப்புகள் (Implications of Psycholoinguistics Findings)

தாய்மொழி பேசுவோர் மொழியின் உருபனியல் அமைப்பின் உள்ளுறை அறிவை (internal knowledge) கொண்டிருக்கிறார்கள் மற்றும் அவர்கள் உருபனியல் கூறிடல் வழி சொற்களஞ்சியத்தை அணுகுகிறார்கள் என்ற உண்மை உளவியல் சான்றுகளால் நிறுவப்பட்டுள்ளது (Caramazza 1988; Taft & Forster 1975; Tyler & Nagy 1990; Holmes & O'

Regan 1992; Laudanna et al 1992, Marslen Wilson 1994) கோர்வை வருடல் மாதிரிகளைவிட (String-Scan Models) வகைப்பாடு-பிளவு மாதிரிகளுக்குக் (Category-Split Models) கூடுதல் முக்கியத்துவம் தரப்படுகிறது என்பதில் ஐயமில்லை.

ஃபாஸ்டரின் மாதிரியை (Foster's Model) சொல்தகவல் மையத்தின் (Hash-table representation) சிறு அட்டவணை உருப்படுத்தமாக (Hash-table representation) கணிப்பொறியியல் அடிப்படையில் பார்க்கலாம். முதல் சில எழுத்துக்களின் அடிப்படையில் உருவாக்கப்பட ஹாஷ் குறியங்கள் (Hash codes) முதன்மைச் சொற்களஞ்சியத்திற்குக் (Main lexicon) கொண்டு செல்லும் சொல்லை அதன் உருப்படுத்தற்கு முன்னரே வருவதுரைக்கும் மாற்ற வலைப்பின்னல் மாதிரியைப் பற்றிய பாஸ்டரின் அனுமானம் கருதப்பட்ட சொல் தூண்டத் தவறிய பதில்செயல் காலத்தைப் பொறுத்தவரையில் தவறாகப் போவதாகத் தோன்றுகின்றது. சொற்கள் உளச்சொற்களஞ்சியத்தில் நிகழ்வெண் தரத்தால் நிரல்படுத்தப்பட்டுள்ளது என்பது உள்ளுணர்வு சார்ந்ததாகும். இம்மாதிரியான கருத்துருவை முழுவதுமாக ஏற்றுக்கொள்ள இயலாதகையால் இதன் அடிப்படையிலான கணிப்பொறி மாதிரியை உருவாக்குவது சாத்தியமில்லாது வரும். இதே வழியில் தானியக்க தேடல் மாதிரியும் பரவலாகப் பயன்படுத்தவில்லை என்று தோன்றுகிறது.

4.3.1.4 கணிப்பொறி வழி ஆய்வுக்கான பொருத்தமான மாதிரி

ஸ்ப்ரோட் பின்வரும் கேள்விகளைக் கேட்கிறார்:

1. ஒழுங்குமுறையில் பயன்படுத்தப்படும் சொல் மீட்பு மாதிரி (Lexical Retrieval Model) உண்மையிலேயே மனிதர்கள் என்ன செய்கிறார்கள் என்பதுடன் ஒத்திருக்கிறதா?

2. கருதத்தக்கத் திறனுள்ள பெரும்பான்மையான கணிப்பொறி மாதிரிகள் உருபனியல் கூறியடலின் கூடுதல் அளவையோ குறைந்த அளவையோ உட்படுத்துகிறது. வழிமுறை வரைவுகளும் (Algorithms) உருப்படுத்தங்களும் மனிதன் செய்யும் முறையின் நியாயமான மாதிரிகளைப் பயன்படுத்துகிறதா?

உருபனியல் ஆய்வில் உளவியல் உண்மை உணர்வுபூர்வமான வழியில் இயற்கை மொழியின் பகுப்பாய்வில் நம்பிக்கை வைக்கும் ஒழுங்கமைப்புக்கு முன்தேவையாகும். மனச் சொற்களஞ்சியத்திலிருந்து சொற்களை அணுகுவது முன்-சொல் சார்ந்ததாகவோ (pre-lexical (முறைமுகமானது) பின்-சொல் சார்ந்ததாகவோ (post-lexical (நேரடியானது) இருக்கலாம் முன்-சொல் அணுகல் சொற்களஞ்சியத்திற்குள் பதிவின் இடம் வரை உள்ளீட்டின் பகுப்பாய்வை உட்படுத்தும். அதாவது சொற்கள் மனச்சொற்களஞ்சியத்திலிருந்து தெரிந்து கொள்வதற்கோ திரும்பப் பெறுவதற்கோ முன்னர் அவற்றின் உறுப்பு உருபன்களாகப் பிரிக்கப்படும். பின்-சொல் பகுப்பாய்வு சொல் பதிவு அணுகப்பட்டபின்னர் நிகழ்கிறது. அதன் ஒலியனியல் மற்றும் எழுத்தியல் வடிவக் குறிப்பீடுகள் ஆய்வுக்காக ஒரே நேரத்தில் கிடைக்கும்.

டாஃப்ட் மற்றும் ஃபாஸ்டர் (Taft and Forster) மாதிரி மீதான ஆய்வுகள் எமோரே (Emmorey, 1998) என்பவரால் வெற்றியடையாதது என்று கருதப்பட்டாலும் பரிசோதனைகள் உண்மையான பகுதிகளையும் ஒட்டுகளையும் போலி அடிகள் மற்றும் போலி ஒட்டுகளை விட எளிதில் இல்லாத சொற்களாகத் (non-words) தள்ளிவைக்கலாம் என்று காட்டின. அனுமானங்கள் எப்படியிருந்தாலும் திரிபு வகுப்பு உருபனியல் கூறிடலின் செயற்பாங்கை நிர்ணயிக்கிறிது என்ற கூற்றுதான் இங்கு பலன் தருவதாகும். மேக்வீன் மற்றும் கட்லர் (McQueen & Cutler) பின்வருமாறு கூறுகின்றனர்: உருபனியல் அடிப்படையில் அமைப்பாக்கம் செய்யப்பட்ட சொல் உருப்படுத்தங்களின் அணுகல் வெளிப்படையாக உருபனியல் அமைப்பை உட்படுத்தத் தேவையில்லை; ஆக்க வடிவுகளின் கூறிடல் முழுச்சொல் அணுகல் நெறிமுறை தோல்வியுறும் போது இருக்கும் விருப்பு நெறிமுறை ஆகும்.

பரிசோதனைகள், ஆங்கில மொழிக்குச் சொற்களின் பகுப்பாய்வு முன்னொட்டுகளை ஒக்கும் அலகுகளின் பரித்தெடுக்கையில் தொடங்கி பகுதிகளுக்கு முன்னேறிப் பின்னர் பின்னொட்டுகளுக்குச் செல்கின்றது என்று நிறுவுகின்றது. இம்மாதிரிகளில் ஒரு பதிவுக்கான ஒத்தறி அணுகல் நேரம் தர்க்கத்திற்கு விஷயமாகும். உருபனியலின் எக்கணினி உருப்படுத்தமும் மனித மனதில் இயற்கை மொழி ஆய்வின் இயல்பை விவரிக்க உளவியல் சான்றின் சில நோக்குகளைச் சார்ந்திருக்கத் தேவையாகும். மாதிரிகளின் பயன்களையும் குறைபாடுகளையும் மனதில் வைத்துக்கொண்டு சொல்லும் வரிசைமுறையும் (IA) அல்லது சொல்லும் செயற்பாங்கும் (IP) அல்லது இரண்டின் பண்புக் கூறுகளும் இணைந்த மாதிரியை அடிப்படையாகக் கொண்ட உருபனியல் ஆய்வைப் போதுமான அளவுக்குக் கணிக்கும் எந்த வகைப்பாடு-பிளவு மாதிரியையும் (Category-split model) பயன்படுத்துவதற்குச் சாதகமானதாகக் கொள்ளலாம். எந்த ஒட்டு எந்தப் பகுதியுடன் சேரும் என்று விவரணை செய்யும் கட்டுப்பாடுகளுடன் பகுதிகள் மற்றும் ஒட்டுகளின் பட்டியலைத் தனித்தன்மையாக ஆம்பிள் உட்படுத்தும். இவை கிம்மோ மாதிரியில் தொடர்ச்சியான அமைப்பொழுங்குகளாகப் பயன்படுத்தப்படும். இது ஒலியனியல் மற்றும் உருபனியல் இவற்றின் முக்கியமான விதிகளுடன் அடிப்படை உருபன் வடிவங்களின் சேமிப்பை முற்கோள் செய்யும். பெர்ல் (PERL) வழியமைப்பு மொழியில் உருவாக்கப்பட்ட பகுப்பாய்வி எளிமை காரணமாக உருபனியல் அடுக்குகள் அடிப்படையில் அமைந்த பகுப்பாய்வு இயக்க நுட்பத்துடன் சொல்லும் வரிசை முறையும் மற்றும் சொல்லும் செயற்பாங்கும் என்ற இரண்டு மாதிரிகளின் பண்புக்கூறுகளை உள்ளடக்கிய மாதிரியாகும் (Vaishnavi Ramaswamy, 2003:48).

4.3.2. தமிழ் உருபனியல் (Tamil Morphology)

தமிழ் உருபனியல் முழுவதுமான/நிறைவான பிணைப்பு உருபனியல் (Concatenative Morphology) எனலாம். இதன்படி உருபுகள் எல்லாம் தொடர்ச்சியாக ஒட்டப்பட்டுள்ளன. தமிழ் பிணைப்பு உருபனியல் பின்னொட்டுகளாக பகுதிக்குப் பின்னர் சேர்க்கப்பட்டுள்ளது. இதைப் பின்வருமாறு உருப்படுத்தம் செய்யலாம்.

பகுதி (+ ஒட்டு)^{n}

இதில் n பின்னொட்டுகள் ஒன்றோ பல தடவையோ வருவதைக் குறிப்பிடும். உருபனியல் மற்றும் தொடரியல் அடிப்படையில் தமிழ்ச் சொற்களை பின்வருமாறு எட்டு சொல் வகுப்புகளாகப் பகுக்கலாம்.

1. பெயர்கள் (Nouns)
2. வினைகள் (Verb)
3. பின்னுருபுகள் (Post Position)
4. பெயரடைகள் (Adjectives)
5. வினையடைகள் (Adverbs)
6. அளவை அடைகள் (ழுஇடிஇவகைநைசள)
7. அடைகொளி அடைகள் (Determiners)
8. இணைப்பிகள் (Conjanction)

4.3.2.1.பெயரும் திரிபும் (Nouns and Inflection)

பெயர்கள் வேற்றுமைக்கும் எண்ணிற்கும் திரிபுறும் எனப் பார்த்தோம். திரிபுற்ற ஒரு பெயர்ச்சொல் பெயர்ப்பகுதி, எண் மற்றும் வேற்றுமை என்ற மூன்று உருபன்களின் ஒலியனியல் உட்படுத்தம் ஆகும். இதைப் பின்வருமாறு வெளிப்படுத்தலாம்:

பெயர் (+ எண்) (+ வேற்றுமை)

ஒலியனியல் நிலையில் திரிபுற்ற பெயர் வடிவம் மேற்சொன்ன மூன்று உருபன்களையும் மெய்ப்படுத்தும் செய்யும் மூன்று உருபுகளை மட்டும் எப்பொழுதும் கொண்டிருப்பதில்லை. ஏனென்றால் பெயர்த் திரிபுன் போது இலக்கண உருபன்களை மெய்ப்படுத்தம் செய்யும் எண் மற்றும் வேற்றுமைப் பின்னொட்டுகள் தவிர சொல் அல்லது இலக்கணச் செயல்பாடில்லாத சில ஒலியனியல் ஒட்டுக்களும் வரலாம். இவ்வொலியனியல் பின்னொட்டுகள் வெற்றுருபன்கள் (empty morpheme) எனப்படும். அதாவது அவை ஒரு உருபனை உருப்படுத்தம் செய்யாது. இலக்கண நூல்களில் இவ்வுருபனியல் பின்னொட்டுகள் சாரியைகள் (inflectional increments) எனப்படும். கட்டாய மற்றும் விருப்பு ஒலியனியல் பின்னொட்டுகள்/சாரியைகள் இவற்றுக்கிடையில் வேறுபாட்டைக் கருதலாம். நாம் கட்டாயத் திரிபு சாரியைக்ளைத் திரிபு (oblique) சாரியைகள் அல்லது திரிபு ஒட்டுகள் என்றும் கூறலாம். விருப்புத் திரிபு சாரியைகளை ஒலி நிரவல் (euphonic) சாரியை அல்லது ஒலி நிரவல் பின்னொட்டுகள் என்றும் குறிப்பிடலாம்.

உள்ளுறை அமைப்பில் இரு இலக்கண உருபன்கள் (ஒன்று எண் ஆகவும் மற்றொன்று வேற்றுமையாகவும் வரும்) பெயர் உருபனுடன் வரும். ஒலியனியல் நிலையில் நான்கு வகையான உருபுகள் (பின்னொட்டுகள்) பெயர்ப் பகுதியுடன் சேர்ந்து வரும். அவை பின்வருவனவாகும்.

1. பன்மைப் பின்னொட்டு
2. திரிபு பின்னொட்டு (சாரியை)
3. ஒலிநிரவல் பின்னொட்டு (சாரியை)
4. வேற்றுமைப் பின்னொட்டு

இந்நான்கு பின்னொட்டுகளும் பெயர்ப்பகுதியுடன் பின்வருமாறு வரும்:

1. பெயர் எண்ணுக்காகத் திரிபுறும்போது பன்மைப் பின்னொட்டு பெயர்ப் பகுதியில் சேர்க்கப்படும்.

   பெயர்ப் பகுதி + பன்மைப் பின்னொட்டு

   எ.கா.

   வீடு- கள், எலி- கள்

2. பெயர் வேற்றுமைக்காகத் திரிபுறும்போது வேற்றுமைப் பின்னொட்டு பெயர்ப் பகுதியுடன் நேரிடையகவோ திரிபுப் பகுதியுடனோ சேர்க்கப்படும்.

   பெயரின் திரிபுப் பகுதி பெயருடன் திரிபு ஒட்டைச் சேர்ப்பதனாலோ நிலை மொழியின் இறுதி மெய்யொலியை இரட்டுவதாலோ உருவாக்கப்படும். ஒலி நிரவல் சாரியை *இன்* (சில நேர்வுகளில் *–அன்*) பெயர் அல்லது திரிபுப் பகுதிக்கும் வேற்றுமைப் பின்னொட்டுக்கும் இடையில் விருப்பாக வரும். இவ்வுருபன் அமைப்பொழுங்கைப் (morphotactics) பின்வருமாறு தரலாம்:

   பெயர்ப்பகுதி / திரிபுப்பகுதி + (ஒலிநிரவல் சாரியை) + வேற்றுமைப் பின்னொட்டு

   வீடு- கள்(-இன்)- ஐ > வீடுகளை/வீடுகளினை

   கை-கள்(-இன்)- ஐ > கைகளை/கைகளினை

பெயர் அல்லது திரிபுப் பகுதியுடன் பன்மை, வேற்றுமை மற்றும் ஒலியன் ஒட்டுகளைச் சேர்க்கும்போது பல உருபொலியனியல் விதிகள் செயல்படும். எடுத்துக்காட்டாக இடைவெளியைத் (hiatus) தவிர்க்க *வ்* அல்லது *ய்* என்ற ஒலியன் வடிவு உடன்படுத்தியாகச் (glide) சொருகப்படும். இதைப்பற்றி பின்னால் விரிவாகக் கூறப்பட்டுள்ளது.

4.3.2.2 பெயர்ப் பகுதி

பெயர்ப்பகுதி என்பது திரிபுப் பின்னொட்டுகள் சேர்க்கப்படாத பெயர் வடிவமாகும். இவ்வடிவில்தான் பெயர்கள் அகராதியில் பட்டியலிடப் பட்டுள்ளன. இவை எழுவாய் வடிவிலும் வரும். பெயர்ப்பகுதி தனியாகவோ கலப்பாகவோ வரலாம். பெயர்ப்பகுதி ஒரு பெயர் வேரைக் கொண்டிருக்கும் (எ.கா. *கல்*); ஒரு கலவைப் பெயர்ப்பகுதி ஒரு வேரையும் (எ.கா. ஒரு வினைப் பகுதியும்) ஒரு ஆக்க ஒட்டையும் (எ.கா. பெயராக்கப் பின்னொட்டு *ப்பு*) கொண்டிருக்கலாம்.

[பெ[வி படி] -ப்பு]

அல்லது அது ஒரு வேரையும் பால் பின்னொட்டையும் கொண்டிருக்கலாம்.

மாண-வ்-அன்>மாணவன், மாண-வ்+இ > மாணவி, மாண-வ்-அர்> மாணவர்

தலை-வ்-அன்> தலைவன், தலை-வ்-இ> தலைவி, தலை-வ்-அர் > தலைவர்

இதில் *-அன்* என்பது ஆண்பால் பின்னொட்டாகும். *-இ* என்பது பெண்பால் பின்னொட்டாகும். *-அர்* என்பது பலர்பால் (உயர்வுப் பன்மை) பின்னொட்டாகும். பல பெயர்ப்பகுதிகள் *அ* என்பதில் முடியும் வேரையும் பகுதியை உருவாக்கும் (stem forming) பின்னொட்டான *ம்* என்பதையும் கொண்டிருக்கும்.

மர-ம் > மரம், பழ-ம் > பழம், நில-ம்> நிலம், திட்ட-ம்> திட்டம்

4.3.2.3 திரிபுப் பகுதி

சில பெயர்கள் பெயர்ப்பகுதியுடன் திரிபுப்பகுதி (oblique stem) என்று அழைக்கப்படும் திரிபுப் பெயர் வடிவத்தையும் கொண்டிருக்கும். சில திரிபுப் பகுதியைக் கொண்டிருக்காது. திரிபுப்பகுதிக்கு நான்கு மடங்கு செயல்பாடும் சூழல் வருகையும் இருக்கும். மேற்சொன்ன எடுத்துக்காட்டில் காட்டியுள்ளது போல் முதலாவது, இப்பெயர் வடிவத்துடன் தான் வேற்றுமைப் பினொட்டுகள் சேர்க்கப்படுகின்றன; இரண்டாவது பெயரின் வடிவம்தான் பெயர் ஒரு அடையாகக் கிழமைச் செயல்பாட்டுடன் தலைமைப் பெயருக்கு முன்னால் வரும் வடிவமாகும்; மூன்றாவது இப்பெயர் வடிவம்தான் சுட்டுப்பெயர் கட்டுமானத்தில் முதல் பெயர்த் தனிமமாக வரும் வடிவமாகும். நான்காவது இப்பெயர் வடிவம்தான் பல பின்னுருபுகளுக்கு முன்வரும் வடிவமாகும். திரிபுப்பகுதியாக வரும். பெயர்கள் திரிபுப் பகுதியாக வருமா வராதா என்பதன் அடிப்படையிலும் திரிபு அடியாக வந்தால் அவற்றின் ஒலியனியல் வடிவம் அடிப்படியிலும் பெயர்களை நான்கு வகுப்புகளாக வேறுபடுத்திக் காட்டலாம்.

1) *த்து* என்ற திரிபு ஒட்டு உடைய திரிபுப் பகுதி
2) *அற்று* என்ற திரிபு ஒட்டுடைய திரிபுப் பகுதி
3) மெய்யொலிகள் இரட்டிப்பால் உருவாக்கப்படும் திரிபுப் பகுதி
4) மாற்றுப் பெயர்களின் திரிபுப் பகுதி

4.3.2.3.1. *த்து* என்ற திரிபு ஒட்டு உடைய திரிபுப் பகுதி

இவ்வகுப்பில் வரும் பெயர்கள் *அம்* என்ற இறுதி அசையைக் கொண்டிருக்கும் (எ.கா. மரம், நிலம், பழம்). இப்பெயர்கள் அவற்றின் இறுதி ஒலியனான *ம்* என்பதை *த்து* என்ற திரிபு ஒட்டால் இடம்பெயர்த்துத் திரிபுப் பகுதியை உருவாக்கும்.

மர-த்து > மரத்து, பழ-த்து>பழத்து, நில-த்து>நிலத்து, தோட்ட- த்து>தோட்டத்து

இத்திரிபுப்பகுதியின் வருகையின் நான்கு சூழல்கள் கீழே எடுத்துக் காட்டப்பட்டுள்ளன.

1. திரிபுப் பகுதி + வேற்றுமை ஒட்டு

   தோட்ட-த்த்-இல் > தோட்டத்தில்

2. திரிபுப்பகுதி + தலைமைப் பெயர்

   தோட்ட-த்து-க் கதவு > தோட்டத்துக் கதவு

3. திரிபுப் பகுதி + பெயர் (பெயர்-பெயர் கூட்டில்)

   தோட்ட- த்து- ப் பூ > தோட்டத்துப் பூ

4. திரிபுப் பகுதி- பின்னுருபு

   தோட்ட-த்து-ப் பக்கம் > தோட்டத்துப் பக்கம்

திரிபு வடிவுகளைத் தொடர்ந்து வரும் பெயர்களின் அல்லது பின்னுருபுகளின் தொடக்க அடைப்பு ஒலிகளாக *க், ச், த், ப்* என்பனவற்றின் அடிப்படையில் முன்னர் வரும் திரிபுப் பகுதி சம்பந்தப்பட்ட அடைப்பொலிக்கு இரட்டித்த ஒலியைக் கொண்டிருப்பதை உணராலாம்.

4.3.2.3.2. *அற்று* என்ற திரிபு ஒட்டுடைய திரிபுப் பகுதி

இவ்வகுப்பு ஐந்து பெயர்களைக் (மாற்றுப் பெயர்களை) கொண்டிருக்கும். இவை ஒரே திரிபு ஒட்டை ஏற்கும் என்பதன் அடிப்படையில் பொதுவான பண்பைக் கொண்டிருக்கின்றன. இவை பல, சில,அவை மற்றும் எல்லாம் என்பனவாகும். இப்பெயர்கள் அவற்றின் திரிபுப் பகுதியை *அற்று* என்ற திரிபு ஒட்டு சேர்க்கப்படும்போது உருவாக்கும். எல்லாம் என்பதில் *ம்* என்பது நீக்கப்படும். *இவை, அவை* என்பனவற்றில் *ஐ* என்பது *அ* என சுருக்கப்படும்.

பல-வ்-அற்று > பலவற்று

அ-வ்-அற்று > அவற்று

எல்லா-வ்-அற்று > எல்லாவற்று

இந்த வகையில் படும் திரிபுப் பகுதி வரும் சூழல்கள் கீழே பட்டியலிடப்பட்டுள்ளன.

1. திரிபுப் பகுதி + வேற்றுமை ஒட்டு

   பல-வ்-அற்-ஓடு > பலவற்றோடு

2. திரிபுப் பகுதி + தலைமைப் பெயர்

பல- வ்- அற்-இன் விலை > பலவற்றின் விலை

3. திரிபுப் பகுதி + முன்னுருபுத் தொடர்

பல-வ்- ற்-இன் பக்கம் > பலவற்றின் பக்கம்

இவ்வகையிற்படும் பெயர்களுடன் (மாற்றுப்பெயர்களுடன்) *இன்* என்ற ஒலிநிரவல் சாரியை கட்டாயமாக அற்று என்ற திரிபு ஒட்டுக்குப் பின்னால் அவை பெயர் விசேடணமாக வரும்போதொ அல்லது பின்னுருபால் தொடரப்படும்போதோ வரும்.

4.3.2.3.3 மெய்யொலிகள் இரட்டிப்பால் உருவாக்கப்படும் திரிபுப் பகுதி

டு அல்லது *று* என்ற அசைகளால் இறுதியுறும் எல்லாப் பெயர்களும் இவ்வகுப்பில் அடங்கும் (எ.கா. *வீடு, நாடு, ஆறு, கிணறு*). இவ்வகுப்பைச் சார்ந்த பெயர்கள் இறுதி அசையின் மெய்யெழுத்துக்கள் இரட்ட திரிபு அடியை உருவாக்கும்.

வீடு > வீட்டு, நாடு > நாட்டு, ஆறு > ஆற்று, கிணறு > கிணற்று

இவ்வகையிற்படும் திரிபு அடியின் வருகைச் சூழல் கீழே பட்டியலிடப்பட்டுள்ளன.

1. திரிபுப் பகுதி + வேற்றுமை ஒட்டு

வீட்- ட்-இல்> வீட்டில்

கிணற்- ற்- இல்>கிணற்றில்

2. திரிபுப் பகுதி + தலைமைப் பெயர்

வீட்-டு-க் கதவு>வீட்டுக் கதவு

கிணற்- று-ச் சுவர்>

3. திரிபுப் பகுதி + பெயர்

வீட்-டு வேலை > வீட்டுவேலை

கிணற்-று-த் தண்ணீர் > கிணற்றுத் தண்ணிர்

4. திரிபுப் பகுதி + முன்னுருபுத் தொடர்

வீட்-டு-ப் பக்கம் >வீட்டுப் பக்கம்

கிணற்-று-ப் பக்கம் >கிணற்றுப் பக்கம்

மேற்சொன்ன மூன்று வகுப்பில் படாத பெயர்ச் சொற்கள் யாவும் பெயரின் நான்காவது வகுப்படிச் சாரும். இவைகள்தாம் பெயர்களில் பெரும்பான்மையாகும் பிற வகுப்பைச் சார்ந்த பெயர்கள் அவற்றின் திரிபுப் பகுதியாக வருமிடங்களில் இவை அவற்றின் பெயர்ப் பகுதியாக வரும்.

1. திரிபுப் பகுதி + வேற்றுமை ஒட்டு

தலை-ய்-இல் > தலையில்

2. திரிபுப் பகுதி + தலைமைப் பெயர்

   தலை மயிர்

3. திரிபுப் பகுதி + பெயர்

   தலை வலி

4. திரிபுப் பகுதி + பின்னுருபு

   தலை மேல்

4.3.2.3.4. மாற்றுப் பெயர்களின் திரிபுப் பகுதி

மாற்றுப் பெயர்கள் பெயர்ச்சொல்லின் வகைப்பாட்டின் ஒரு அடைப்பட்ட துணைக் குழுமமாகும். மாற்றுப் பெயர்கள் பெயர்களின் துணைக்குழுமமாகக் குழுமப்படுவதின் காரணம் பெரும்பாலும் தொடரியல் சார்ந்ததாகும். எனவே மாற்றுப் பெயர்கள் தொடரியல் வகைப்பாட்டைச் சார்ந்ததாகக் கருதப்படிகின்றன. ஐந்து வகைப்படும் மாற்றுப் பெயர்களின் ஒரு வகைக்கு அதாவது மூவிட மாற்றுப் பெயர்களுக்குத் திரிபுப்பகுதி இருக்கின்றது. முதலாவது, இரண்டாவது மற்றும் நான்காவது இடமாற்றுப் பெயர்களும், மூன்றாவது இடப்பன்மை அஃறிணை மாற்றுப் பெயர்களுக்கும் திரிபுப்பகுதி இருக்கின்றது. முதலாவது, இரண்டாவது மற்றும் நான்காவது மாற்றுப் பெயர்கள் அவற்றின் திரிபுப் பகுதியை அவற்றின் மாற்றுப்பெயர்கள் வடிவின் ஒலியனியல் மாற்றத்தால் உருவாக்கின்றது. படர்க்கை இடப் பன்மை அஃறிணை மாற்றுப்பெயர்களுக்கு அற்று திரிபு ஒட்டுபெற்ற திரிபுப்பகுதி இருக்கின்றது. பின்வரும் அட்டவணை மாற்றுப் பெயர்களையும் அவற்றின் திரிபு வடிவங்களையும் காட்டும்.

இடம்	**மாற்றுப்பெயர்**	**திரிபு வடிவம்**
தன்மை ஒருமை	நான்	என் (ன்), என
முன்னிலை ஒருமை	நீ	உன் (ன்), உன்
படர்க்கை ஒருமை ஆண்பால்	இவன்/அவன்	இவன்/அவன்
படர்க்கை ஒருமை பெண்பால்	இவள்/அவள்	இவள்/அவள்
படர்க்கை ஒருமை அஃறிணை	இது/அது	இது/அது
படர்க்கை ஒருமை ஆண்பால் / பெண்பால்	இவர்/அவர்	இவர்/அவர்
தற்சுட்டு ஒருமை	தான்	தன்/தன
தன்மைப் பன்மை	நாங்கள்/நாம்	எங்கள்/ நம்(ம்), நம
முன்னிலைப் பன்மை	நீங்கள்	உங்கள்
படர்க்கைப் பன்மை ஆண்பால்/ பெண்பால்	இவர்கள்/அவர்கள்	இவர்கள்/அவர்கள்

படர்க்கை பன்மை/ அஃறிணை	இவை/அவை இவைகள்/அவைகள்	இவற்று/அவற்று இவைகள்/அவைகள்
தற்சுட்டுப் பன்மை	தாங்கள்	தங்கள்

மேற்சொன்ன அட்டவணைப்படி *நான்* என்ற தன்மைப் பெயருக்கு மூன்று திரிபுப் பகுதி வடிவங்கள் இருக்கின்றன: *என்*, *என்ன்*, *என. என* நான்காவது வேற்றுமைப் பின்னொட்டுக்கு முன்னும், *என்ன்* என்பது *இடம்* என்ற பின்னுருபுக்கு முன்னும், *என்* தலைமைப் பெயருக்கு முன்னும் பிற பின்னுருபுகளுக்கு முன்னும் வரும்.

1. திரிபுப் பகுதி + வேற்றுமை ஒட்டு

   என-க்கு > எனக்கு, என்ன்-இடம் > என்னிடம்

2. திரிபுப் பகுதி + தலைமைப் பெயர்

   என் வீடு

3. திரிபுப் பகுதி + பின்னுருபு

   என் பக்கம்

4.3.2.3.5. ஒலி நிரவல் சாரியைகள் *இன், அன்*

முன்னர் *த்து, அற்று* என்ற திரிபுச் சாரியைகளைப் பற்றிப் பார்த்தோம். இவை தவிர *இன், அன்* என்ற ஒலியனியல் பின்னொட்டுக்கள் வடிவத்திலிருக்கும் விருப்புத் திரிபுச் சாரியைகளும் இருக்கின்றன. இவை ஒலி நிரவல் முன்னொட்டுக்கள் என்று அழைக்கப்பெறும். *இன்* என்பது எல்லாப் பெயர்களுடன் மட்டும் வரும். *அன்* என்பது *இது* மற்றும் *அது* என்ற மாற்றுப்பெயர்களுடன் மட்டும் வரும். இவ்வொலி நிரவல் ஒட்டுக்களான *இன்* என்பதும் *அன்* என்பதும் வரும் சூழல்கள் கீழே தரப்பட்டுள்ளன:

1. வேற்றுமைக்காகத் திரிபுற்ற பெயர்களில் ஒலிநிரவல் பின்னொட்டு பெயர்ப்பகுதிக்கும் அல்லது திரிபுப் பகுதிக்கும் அல்லது பன்மை ஒட்டுக்கும் வேற்றுமைப் பின்னொட்டுக்கும் இடையில் வரும்.

   கை(-இன்)-ஆல் > கையால்/கையினால்

   அத்(-அன்)-ஆல் > அதால்/அதனால்

   யுத்த-த்த்(-இன்)-ஆல் > யுத்தத்தால்/யுத்தத்தினால்

   கற்-கள்(-இன்)-ஆல் >கற்களால்/கற்களினால்

2. அடையாக வரும் பெயர்களுடன் உடைமை வேற்றுமைச் செயல்பாட்டில் தன்மைப் பெயருக்கு முன் வரும்.

   கை(-ய்-இன்) புத்தகம் > கையின் புத்தகம்

மர(-த்த்-இன்) கிளை > மரத்தின் கிளை

பையன்-கள்(-இன்) சட்டை > பையன்களின் சட்டை

3. ஒரு கிளமைப் பின்னொட்டுக்கு முன் விருப்பாக, சில பெயர்களுடன் கட்டாயமாகப் பெயர்ப்பகுதிக்கு அல்லது திரிபுப்பகுதிக்கு அல்லது பன்மை ஒட்டுக்குப் பின் வரும்.

சாவி(ய்-இன்) மூலம் > சாவியின் மூலம்

அத்(-அன்) மூலம் > அதன் மூலம்

மர(-த்த்(-இன்) மூலம் > மரத்தின் மூலம்

ஆணி-கள்(-இன்) மூலம் >ஆணிகளின் மூலம்

4.3.2.4. எண் குறிப்பு

தமிழில் ஒருமை, பன்மை என்ற இரண்டு எண்கள் இருக்கின்றன. பன்மை ஒட்டு *-கள்/-க்கள்* என்பதால் உருப-ஒலியனியல் அடிப்படையில் குறிப்பீடு செய்யப்பட்டுள்ளது. முன்னர் கூறியபடி பன்மை ஒட்டு பெயர்ப்பகுதியுடன் சேர்க்கப்படுகிறது; பெயரின் திரிபு அடியுடன் சேர்க்கப்படுவதில்லை. பன்மை ஒட்டிற்கு *கள்*, *க்கள்* என்ற இருமாற்றுருபுகள் இருக்கின்றன. அவற்றின் வருகை பின்வருமாறு அமையும்: *க்கள்* நெட்டுயிர் இறுதியில் வரும் எல்லாப் பெயர்களுடனும் இரண்டு குற்றுயிர் உள்ள அசைகளைக் கொண்ட உ என்ற குற்றுயிரால் இறுதியுறும் பெயர்களுடனும் சேர்ந்து வரும்; பிற இடங்களில் *கள்* வரும்.

ஒருமைப் பெயர்	-	பன்மைப் பெயர்	ஒருமைப்பெயர்	-	பன்மைப் பெயர்
பூ	-	பூ- க்கள்	கண்	-	கண்- கள்
ஈ	-	ஈ- க்கள்	மரம்	-	மரங்- கள்
பசு	-	பசு- க்கள்	வீடு	-	வீடு-கள்
தெரு	-	தெரு- க்கள்	கல்	-	கற்- கள்

4.3.2.5. வேற்றுமைக் குறியீடு/உருபு

பெயர் அல்லது பெயர்த்தொடருக்கும் வினைக்கும் அல்லது இரண்டு பெயருகளுக்கும் அல்லது பெயர்த்தொடர்களுக்கும் இடையில் வரும் தொடரியல் மற்றும் பொருண்மையியல் உறவுகளை வேற்றுமை வெளிப்படுத்தும். வினைப் பயனிலைகள் ஒரு வாக்கியத்தில் கட்டாய மற்றும் விருப்புப் பங்கெடுப்பாளர்களை எடுக்கும் (எ.கா. எழுவாய்ப் பெயர்த் தொடர்கள், செயப்படுபொருள், பெயர்த்தொடர்கள், வினையடைகள்). இவ்வாறு வேற்றுமைக் குறியீடுகள் (உருபுகள்) பெயர்த்தொடருக்கும் வினைப் பயனிலைக்கும் இடையிலுள்ள தொடரியல் மற்றும் பொருண்மையியல் உறவை வெளிப்படுத்துகிறது. அதாவது அவை ஒரு பங்கெடுப்பாளர் பெயர்த் தொடரின் இலக்கணச் செயன்மையைக் காட்டும் (எ.கா. எழுவாய் செயல்படு பொருள்). மேலும்

அவை பயனிலையாக்கத்தில் பங்கெடுப்பாளர் பெயர்த்தொடரின் பொருண்மையியல் பங்களிப்பை வெளிப்படுத்தும் (எ.கா. ஒரு செயலின் அல்லது நிகழ்வின் காரணம், நிகழ்வு அல்லது நிலையின் இடம், இயக்கத்தின் மூலம் அல்லது இலக்கு). உடைமை வேற்றுமை இரண்டு பெயர்த்தொடருக்கு இடையில் உள்ள உறவை வெளிப்படுத்தும்.

உருபனியல், தொடரியல் அடிப்படையில் வேற்றுமைக் குறியீடு இரு வழிகளில் உருப்படுத்தம் செய்யப்படுகிறது.

1. வேற்றுமைப் பின்னொட்டுகளால்
2. பின்னுருபுகளால் (இவற்றில் சில கட்டுண்ட வடிவுகள்)

பின்னொட்டுக்களைப் பொறுத்தவரையில் வேற்றுமைக் குறிப்பு வேற்றுமை பின்னொட்டு மற்றும் பின்னுருபு இவற்றின் ஒன்றுசேர்க்கையால் உருப்படுத்தம் செய்யப்படுகிறது. எடுத்துக்காட்டாகக் கருவியின் பொருணமைப் பங்களிப்பு *ஆல்* என்ற பின்னொட்டாலும் (எ.கா. *சாவியால்*) மூலம் என்ற பின்னுருபாலும் (எ.கா. *சாவி மூலம்*) வெளிப்படுத்தலாம் இறுதியாகக் கருவியின் பொருண்மையியல் பங்களிப்பை ஒரு வேற்றுமைப் பின்னொட்டை ஆளும் பின்னுருபால் வெளிப்படுத்தலாம். கொண்டு என்ற பின்னுருபு இரண்டாம் வேற்றுமை உருவான *ஐ*யை ஆளுகை செய்யும் (எ.கா. *சாவியைக் கொண்டு*). தற்காலத் தமிழில் ஆறு வேற்றுமைப் பின்னொட்டுகள் இருக்கின்றன.

வேற்றுமை	**பின்னொட்டுகள்**
செயப்படுபொருள்	ஐ
கொடைவேற்றுமை	(உ)க்கு
கருவி வேற்றுமை	ஆல்
உடன்நிகழ்ச்சி வேற்றுமை	ஓடு
இட வேற்றுமை	இல்
வேற்றுமை	அது

ஐந்து கட்டுண்ட பின்னுருபுகள் இருக்கின்றன.

கட்டுண்ட வேற்றுமை	**கட்டுண்ட பின்னுருபுகள்**
இடவேற்றுமை	இடம்
உடன்நிகழ்ச்சி வேற்றுமை	உடன்
மூல வேற்றுமை	இருந்து
பயனடை வேற்றுமை	ஆக

மரபு இலக்கணவியலார் வேற்றுமைப் பின்னொட்டுகளையும் *ஆக* தவிர பிற கட்டுண்ட பின்னுருபுகளையும் பல இடப் பின்னுருபுகளையும் வேற்றுமை குறிகளாக எடுத்தாண்டனர். இக்குறிகளின் அடிப்படையில் அவர்கள் எட்டு வேற்றுமைகளை வேறுபடுத்தினர். இவைகள்

அவற்றின் ஒலியனியல் வடிவத்தாலோ வரிசை எண்ணாலோ அடையாளப்படுத்தப்பட்டன. வேற்றுமைக் குறியீட்டால் குறிப்பிடப்படாத விளி வேற்றுமையும் அவர்கள் சேர்த்துக்கொண்டனர். பின்னர் மேனாட்டு இலக்கணத்தார்கள் இலத்தின் அல்லது சமஸ்கிருத இலக்கணக் கலைச்சொற்களைப் பயன்படுத்தி வேற்றுமை குறியின் செயல்பாடு அடிப்படையில் எட்டு வேற்றுமைகளை புலக்குறிப்பு செய்தனர்

1. எழுவாய் வேற்றுமை
2. செயல்பாடு பொருள் வேற்றுமை
3. கொடை வேற்றுமை
4. கருவி வேற்றுமை
5. உடன்நிகழ்ச்சி வேற்றுமை
6. இட வேற்றுமை
7. நீங்கல் வேற்றுமை
8. உடைமை வேற்றுமை
9. விளி வேற்றுமை

வேற்றுமைக் குறிகளின் மேற்சொன்ன புலக்குறிப்புகள் ஒரு வேற்றுமைக் குறியின் பொருண்மையை முழுவதுமாக உருப்படுத்தம் செய்யவில்லை; ஏனென்றால் அவை அவற்றின் செயல்பாடுகளில் ஒன்றைத்தான் வெளிப்படுத்துகின்றது. எடுத்துக்காட்டாகக் கருவி வேற்றுமைப் பின்னொட்டு என்று புலக்குறிப்பு செய்யப்பட்ட *ஆல்* என்ற வேற்றுமைப் பின்னொட்டு கருவிச் செயல்பாட்டை வெளிப்படுத்துவதுடன் பிற கருத்துக்களையும் வெளிப்படுத்துகின்றன.

வேற்றுமைக் குறிகளின் மேற்சொன்ன புலக்குறிப்புகள் ஒரு வேற்றுமைக் குறியின் பொருண்மையை முழுவதுமாக உருப்படுத்தம் செய்யவில்லை; ஏனென்றால் அவை அவற்றின் செயல்பாடுகளில் ஒன்றைத்தான் வெளிப்படுத்துகின்றது. எடுத்துக்காட்டாகக் கருவி வேற்றுமைப் பின்னொட்டு என்று புலக்குறிப்பு செய்யப்பட்ட *ஆல்* என்ற வேற்றுமைப் பின்னொட்டு கருவிச் செயல்பாட்டை வெளிப்படுத்துவதுடன் பிற கருத்துக்களையும் வெளிப்படுத்துகின்றன.

வேற்றுமை அடையாளக் குறியீட்டைப் பொருத்தவரையில் பின்வரும் பாகுபாட்டு அடிப்படையில் விளக்கலாம்:

1. வேற்றுமைப் பின்னொட்டு
2. கட்டுண்ட பின்னுருபுகள்
3. பின்னுருபுகள் (சுதந்திரமான வடிவுகள்)

பின்வரும் அட்டவணை வேற்றுமை உருபுகளையும் அவற்றின் வருகை முறைகளையும் அவற்றின் வெறுபட்ட செயல்பாடுகளையும் தரும்:

வேற்றுமை	**பின்னொட்டு/ பின்னுருபு**	**வருகை முறை**	**பொருள்/ செயல்பாடு**
1. எழுவாய் வேற்றுமை	-Ø		**1. எழுவாய்** எ.கா. *ராமன் வந்தான்* **2. பயனிலை** எ.கா. *குமரன் மாணவன்* **3. எழுவாய் நிரப்பி** எ.கா. குமரன் *மாணவன்* ஆவான் 4. **செயப்படுபொருள் நிரப்பி** எ.கா. குமரன் *ராமனைத் தலைவன்* ஆக்கினான் 5. **செயப்படுபொருள்** எ.கா. குமார் *தண்ணீர்* கேட்டான்
2. செயப்படு பொருள் வேற்றுமை	-ஐ		**1. செயப்படு பொருள்** எ.கா. ராமன் பையனை பார்த்தான்
3. கொடை வேற்றுமை	-க்கு~ -உக்கு~-கு	1. பெயர்களின் மற்றும் மாற்றுப் பெயர்களின் திரிபுப் பகுதியுடன் -க்கு வரும். எ.கா. மரத்து-க்கு பலவற்று-க்கு, ஆற்று-க்கு, என-க்கு 2. இ/ஈ/ஐ/அய் என்று முடியும் பெயரடிகளுடன் க்கு வரும் எ.கா. தம்பிக்கு, ஈக்கு, சட்டைக்கு நாய்க்கு, படிப்புக்கு	**1. அயற்செயப்படு பொருள்** எ.கா. குமார் *அப்பாவுக்கு* ஒரு படத்தைக் காட்டினான் **2. இயக்கத்தின் இலக்கு** எ.கா. குமார் *அப்பாவுக்கு* ஒரு கடிதம் கொடுத்தான் **3. கார்ணம்** எ.கா. குமார் *பணத்துக்குத்தான்* வேலை செய்கிறான் **4. அனுபவிப்பவர்** எ.கா. *குமாருக்கு* ஒரு புத்தகம் இருக்கிறது. **5. காலத்தின் கூறு** எ.கா. குமார் *ஐந்து மணிக்கு* வருவான்

		3. பிற எல்லாப் பெயரடிகளுக்குப் பின் உக்கு வரும். எ.கா. அம்மா-உக்கு, பால்- உக்கு, தெரு-வ்-உக்கு, பையன்-உக்கு, பெண்ண்-உக்கு 4. கு ஒலிநிரவல் சாரியை *இன்*, *அன்* என்பனவற்றைத் தொடர்ந்து வரும் எ.கா. பசு+இன்+கு> பசுவிற்கு, அது+அன்+கு அதற்கு	6. **விகிதம்** எ.கா. மருந்து ஒரு *நாளைக்கு* மூன்று வேளை சாப்பிடு 7. **பகிர்ந்தளிப்புச் செயல்பாடு** எ.கா. *ஆளுக்கு* ஒரு டீ போடு 8. **ஒப்புமை மாதிரி** எ.கா. *குமாருக்கு* இவன் நல்லவன் 9. **குறிப்பிடும் இடம்** எ.கா. *மதராசுக்கு* நூறு மைல் பின்னால்
4. பயனடை வேற்றுமை	-க்கு-ஆக ~ -உக்கு-ஆக ~ -கு-ஆக	1. க்கு-ஆக எ.கா. நாய்-க்கு-ஆக 2. உக்கு-ஆக எ.கா.அவன்-உக்கு-ஆக 3. கு- ஆக எ.கா.பசுவிற்-கு-ஆக	1. **பயனடைவு** எ.கா. ராஜா தன் பையனுக்காகப் பணம் சேர்த்தான் 2. **காரணம்** எ.கா. ராஜா தன் *அப்பாவுக்காகச்* சீக்கிரம் வீட்டுக்கு வாந்தான்
5. கருவி வேற்றுமை	-ஆல்		1. **கருவி** எ.கா. குமார் கத்தியால் பழத்தை வெட்டினான் 2. **வழி** எ.கா. குமார் தன் கண்ட *அனுபவத்தால்* இதைக் துளைக்கக் கற்றுக்கொண்டான். 3. **மூலம்/ உருபொருள்** எ.கா. குமார் *மண்ணால்* இந்தப்

			பந்தைச் செய்தான். 4. **காரணம்-1** எ.கா. *வெயிலால்* குமார் வரவில்லை. 5. **காரணம் -2** எ.கா. *மழையால்* பயிர் நன்றாக வளர்ந்தது
6. உடனிகழ்ச் சி வேற்றுமை	-ஓடு/ -உடன்		1. **உடனிகழ்ச்சி** எ.கா. குமார் தன் *மனைவி-ய் -ஓடு/-உடன்* வந்தான் 2. **முறை** எ.கா. குமார் *அன் ப்-ஓடு/-*உடன் சிரித்தான். 3. **வழி** எ.கா குமார் அப்பாவின் *உதவி-ய்-ஓடு/-*உடன் இதைப் படித்தான். 4. **அதிகப்படி** எ.கா. ராஜா நான்கு *இட்லியோடு* இரண்டு தோசையும் சாப்பிட்டான். 5. **இடம்** எ.கா. குமார் *வீட்டோடு* இருக்கிறான். 6. **திசை** எ.கா. குமார் *தெருவோடு* நடந்தான் 7. **உட்படுதல்** எ.கா. சித்திரை *மாசத்தோடு* தமிழ் வருஷம் ஆரம்பிக்கிறது. 8. **வரையறை**

			எ.கா. இந்த ஊரில் ஒரு *கோயிலோடு* சரி
7. இட வேற்றுமை	-இல்		1. **இடத்தில் இடம்** எ.கா. *குருவி* மரத்தில் உட்காந்திருக்கிறது. 2. **காலத்தின் புள்ளியிடம்** (எதிர்காலத்துடன்) எ.கா. ராஜா *ஒரு வாரத்தில்* வருவான். 3. **கால அளவு** (இறந்த காலத்துடன்) எ.கா. ராஜ ஒருவாரத்தில் இந்த புத்தகத்தைப் படித்தான். 4. **பாங்கு** எ.கா. குமார் தமிழில் பேசினான் 5. **இடைபடு உறவு** எ.கா. அந்த மாம்பழங்களில் சில் மல்கோவா. அந்த மாணவர்களில் சிலர் நன்றாக படிப்பவர்கள்.
	-இடம்		1. **இயக்கத்தின் இலக்கு** எ.கா. ராஜா *அப்பாவிடம்* வந்தான். 2. **கைமாற்றத்தின் மூலம்** எ.கா. ராஜா *மந்திரியிடம்* ஒரு பரிசு வாங்கினான். 3. **உணர்ச்சியின் இலக்கு** எ.கா. ராஜா *அப்பாவிடம்* கோபப்பட்டான்.
8. நீக்கல் வேற்றுமை	-இல்/-இடம்–	1. இட வேற்றுமைக்குத்	1. **இயக்கத்தின் மூலம்**

	இருந்து, இருந்து	திரிபுற்ற பெயரைத் தொடர்ந்து வரும் . எ.கா. மரத்தில் – இருந்து பையனிடம் – இருந்து 2. நேரடியாக இடப்பகுதியுடன் வரும். எ.கா. அங்க்- இருந்து மேல்- இருந்து *-இல்* என்ற இடவேற்றுமைப் பின்னொட்டிற்குத் திரிபுற்ற விலங்கினப் பெயர்களுடன் வருகையிலும் *-இடம்* என்ற இட வேற்றுமைப் பின்னொட்டிற்குத் திரிபு விலங்கினப் பெயர்களுடன் வருகையிலும் *-இருந்து* அதன் செயல்பாடுகளில் (பொருள்களில்) வேறுபடும் (பார்க்க அடுத்தப்பக்கத்தில்)	எ.கா. ராஜா *கோவிலிலிருந்து* வந்தான். 2. **உள்ளத்திலிருந்து பிரிதல்** எ.கா. ராஜா தன் அப்பாவை *நெருப்பிலிருந்து/ ஒரு புலியிடமிருந்து* காப்பாற்றினான். 3. **தேர்வின் மூலம்** எ.கா. ராஜா *மாணவர்களிலிருந்து* ஒரு நல்லவனைத் தேர்ந்தெடுத்தான். 4. **தொடர்ச்சியின் தொடக்கம்** எ.கா. *மந்திரியிலிருந்து* எல்லோரும் லஞ்சம் வாங்குகிறார்.
9. உடைமை வேறுமை	-இன், -உடைய, -அது, பெயர்திரிபு அடி	1. **திரிபு வடிவில் பெயர்** எ.கா. மரத்துக்கிளை 2. **பெயர் + ஒலிநிரவல் – *இன்*** எ.கா. *அரசனின்* முடி 3. **பெயர் + உடைமை வேற்றுமை** *அது*	**பெயரடையாகச் செயல்படுகிறது** எ.கா. பள்ளிக்கூடம் குமாருடையது.

		எ.கா.அரசனது முடி 4. **பெயர் + இடைநிலை சாரியை + உடைமைப் பின்னொட்டு** எ.கா. *நாயினது* வால் 5. **பெயரடைப் பின்னொட்டு உடைய** எ.கா. *நாயினுடைய* வால்	

4.3.2.6. பெயர் அடுக்குகள்

பெயர்கள் வேற்றுமைகளுக்காக திரிபுறுவதைக் கருத்தில் கொண்டு இரண்டு பெயர்களின் வேற்றுமை சார் அடுக்குகள் கீழே தரப்பட்டுள்ளன, ஒன்று அஃறிணைப் கால் என்பதற்காகவும் மற்றொன்று பன்மைப் பெயர் பையன் என்பதற்காகவும் வேற்றுமை அடுக்குகள் தரப்பட்டுள்ளன.

வேற்றுமை	**ஒருமை**	**பன்மை**
எழுவாய்வேற்றுமை	கால்	கால்-கள்
செயப்படுபொருள்வேற்றுமை	கால்-ஐ	கால்-கள்-ஐ
கொடை வேற்றுமை	கால்-உக்கு	கால்-கள்-உக்கு
நன்மை வேற்றுமை	கால்-உக்கு-ஆக	கால்-கள்-உக்கு-ஆக
கருவி வேற்றுமை	கால்-ஆல்	கால்-கள்-ஆல்
உடனிகழ்சி வேற்றுமை	கால்-ஓடு, கால்-உடன்	கால்-கள்-ஓடு, கால்கள்-உடன்
இட வேற்றுமை	கால்-இல்	கால்-கள்-இல்
நீங்கல் வேற்றுமை	கால்-இல்-இருந்து	கால்-கள்-இல்-இருந்து
உடைமை வேற்றுமை	கால்-இன் –அது	கால்-கள்-அது

வேற்றுமை	**ஒருமை**	**பன்மை**
எழுவாய்வேற்றுமை	பையன்	பையன்-கள்
செயப்படுபொருள்வேற்றுமை	பையன்-ஐ	பையன்-கள்-ஐ
கொடை வேற்றுமை	பையன்-உக்கு	பையன்-கள்-உக்கு

நன்மை வேற்றுமை	பையன்-உக்கு-ஆக	பையன்-கள்-உக்கு-ஆக
கருவி வேற்றுமை	பையன்-ஆல்	பையன்-கள்-ஆல்
உடனிகழ்சி வேற்றுமை	பையன்-ஓடு, பையன்-உடன்	பையன்-கள்-ஓடு, பையன்கள்-உடன்
இட வேற்றுமை	பையன்-இடம்	பையன்-கள்-இடம்
நீங்கல் வேற்றுமை	பையன்-இல்-இருந்து	பையன்-கள்-இல்-இருந்து
உடைமை வேற்றுமை	பையன்-இன் –அது	பையன்-கள்-அது

4.3.2.6.1. வேற்றுமை உருபுகளும் சந்தி விதிகளும்

இராசேந்திரன் (Rajendran, 1999) தமது தமிழ் உருபனியல் ஆய்வி உருவாக்கதிற்கு (Tamil Morphological Analyzer) வேண்டி தமிழ்ப் பெயர்ச்சொற்களைச் சந்திவிதிகளைக் கருத்திற்கொண்டு பாகுபாடு செய்கையில் பின்வரும் சந்திவிதிகளைப் பட்டியலிடுகின்றார் (இராசேந்திரன், 1994:47-56). இது பெயர்களை அவற்றின் எண் மற்றும் வேற்றுமைத் திரிபின் அடிப்படையில் வகைபடுத்த உதவும்.

1. *ஆ, ஓ, ஊ* என முடியும் பெயர்களைத் தொடர்ந்து உயிர்த் தொடக்க வேற்றுமைப் பின்னொட்டுகள் (உ.வே.பி.) வரும்போது *வ்* என்ற உடன்படுத்தி இடையில் வரும்.

-ஆ/ஓ/ஊ + உ.வே.பி→-ஆ/ஓ/ஊ- வ்- உ.வே.பி.

எ.கா.

கடா + ஐ > கடாவை

கடா + ஆல் > கடாவால்

கடா + ஓடு >கடாவோடு

கடா + உடன் > கடாவுடன்

கடா + உக்கு >கடாவுக்கு

கடா + இன் > கடாவின்

கடா + இல் > கடாவில்

2. *இ, ஈ, ஐ* என்று முடியும் பெயர்கள் -உக்கு தவிர்த்த பிற வேற்றுமைப் பின்னொட்டுகளை ஏற்கும்போது *ய்* என்ற உடன்படுத்தி ஒலி இடையில் வரும்.

-இ/ஈ/ஐ + உக்கு தவிர்த்த உ.வே.பி. →இ/ஈ/ஐ–ய்-உ.வே.பி

எ.கா.

எலி + ஐ > எலியை

எலி + ஆல் > எலியால்

எலி + ஓடு > எலியோடு

எலி + உடன் > எலியுடன்

எலி + இன் > எலியின்

எலி + இல் > எலியில்

3. *இ, ஈ, ஐ, ய்* என்று முடியும் பெயர்கள் முவரும் வேற்றுமைப் பின்னொட்டு -*உக்கு* என்பதன் உ கெடும்

-இ/ஈ/ஐ/ய் + உக்கு→இ/ஈ/ஐ/ய்-க்கு

எ.கா.

எலி + உக்கு > எலிக்கு

ஈ + உக்கு > ஈக்கு

யானை + உக்கு > யானைக்கு

நெய் + உக்கு > நெய்க்கு

4. உ-வில் முடியும் பெயர்களை அவற்றின் சந்திச் செயல்பாடு அடிப்படையில் நான்கு வகைகளாகப் பிரிக்கலாம்

4.1. (C)VV-ட்-உ என்ற அசை அமைப்பக் கொண்ட சொற்களைத் தொடர்ந்து உயிர்த் தொடக்க வேற்றுமைப் பின்னொட்டுக்கள் வரும்போது உ கெட, ட் இரட்டிக்கும்.

(C)VV ட் – உ + உ.வே.பி →(C)VVட்ட் – உ.வே.பி

எ.கா.

காடு+ஐ > காட்டை

காடு+ஆல் > காட்டால்

காடு+ஓடு > காடோடு

காடு+உடன் > காட்டுடன்

காடு+உக்கு > காட்டுக்கு

காடு+இன் > காட்டின்

காடு+இல் > காட்டில்

4.2. *று*-வில் முடியும் பெயர்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டுக்கள் வரும்போது உ கெட *ற்* இரட்டும்.

-று + உ.வே.பி → ற்ற் – உ.வே.பி.

4.3. (C)VCC-உ அல்லது (C)VVC-உ என்ற அசை அமைப்புடைய பெயர்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டுகள் வருகையில் உ கெடும்

(C)VCC-உ + உ.வே.பி. → CVCC-உ.வே.பி

எ.கா.

வண்டு+ஐ > வண்டை

வண்டு+ஆல் > வண்டால்

வண்டு+ஓடு > வண்டோடு

வண்டு+உடன் > வண்டுடன்

வண்டு+உக்கு > வண்டுக்கு

வண்டு+இன் > வண்டின்

வண்டு+இல் >வண்டில்

காது+ ஐ > காதை

காது+ ஆல் > காதால்

காது+ ஓடு > காதோடோடு

காது+ உடன் > காதுடன்

காது+ உக்கு > காதுக்கு

காது+ இன் > காதின்

காது+ இல் > காதில்

4.4 (C)VCV என்ற அசையமைப்பைக் கொண்டு உ-வில் முடியும் பெயர்களைத் தொடர்ந்து உயிர்த் தொடக்க வேற்றுமை ஒட்டுகள் வரும்போது *வ்* என்ற உடன்படுத்தி இடையில் வரும்.

(C)VC-உ + உ.வே.பி.→ (C)VC-உ-வ்-உ.வே.பி.

பசு+ ஐ > பசுவை

பசு+ஆல் > பசுவால்

பசு+உக்கு > பசுவுக்கு

பசு+ ஓடு > பசுவோடு

பசு+உடன் > பசுவுடன்

பசு+இன் > பசுவின்

பசு +இல் > பசுவில்

5. (C)VC என்ற அசை அமைப்புடைய *ண், ல், ள், ன்* என முடியும் பெயர்களைத் (ஓரசைச் பெயர்களை) தொடர்ந்து உயிர் தொடக்க வேற்றுமைப் பின்னொட்டுகள் வர முறையே *ண், ய், ல், ள், ன்* என்பன இரட்டும். (விதிவிலக்கு *ய்* + *உக்கு* > *ய்க்கு*. எ.கா. *நெய்* + *உக்கு* > *நெய்க்கு*)

(C)Vண்/ய்/ல்/ள்/ன் + உ.வே.பி→ (C)Vண்ண்/ய்ய்/ல்ல்/ள்ள்/ன்ன். உ.வே.பி.

கண்+ ஐ > கண்ணை

கண்+ஆல் >கண்ணால்

கண்+உக்கு > கண்ணுக்கு

கண்+ஓடு > கண்ணோடு

கண்+உடன் > கண்ணுடன்

கண்+இன் > கண்ணின்

கண்+உடைய > கண்ணுடைய

கண்+அது > கண்ணது

கண்+இல் > கண்ணில்

நெய்+ ஐ > நெய்யை

நெய்+ஆல் > நெய்யால்

நெய்+உக்கு > நெய்க்கு

நெய்+ஓடு > நெய்யோடு

நெய்+உடன் > நெய்யுடன்

நெய்+இன் > நெய்யின்

நெய் +உடைய > நெய்யுடைய

நெய்+அது > நெய்யது

நெய்+இல் > நெய்யில்

கல்+ ஐ > கல்லை

கல்+ஆல் > கல்லால்

கல்+உக்கு > கல்லுக்கு

கல்+ஓடு > கல்லோடு

கல்+உடன் > கல்லுடன்

கல்+இன் > கல்லின்

கல்+உடைய > கல்லுடைய

கல்+அது > கல்லது

கல்+இல் > கல்லில்

முள்+ ஐ > முள்ளை

முள்+ஆல் >முள்ளால்

முள்+உக்கு >முள்ளுக்கு

முள்+ஓடு >முள்ளோடு

முள்+உடன் > முள்ளுடன்

முள்+இன் > முள்ளின்

முள் + உடைய > முள்ளுடைய

முள்+அது > முள்ளது

முள்+இல் >முள்ளில்

பொன்+ ஐ > பொன்னை

பொன்+ ஆல் > பொன்னால்

பொன்+உக்கு > பொன்னுக்கு

பொன்+ஓடு > பொன்னோடு

பொன்+உடன் > பொன்னுடன்

பொன்+இன் > பொன்னின்

பொன்+உடைய > பொன்னுடைய

பொன்+அது > பொன்னது

பொன்+இல் > பொன்னில்

6. *அம்* ஈற்றுப் பெயர்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டுக்கள் வரும்போது *ம்* கெட *த்த்* என்ற சாரியை இடையில் வரும்.

-அம் + உ.வே.பி.→-அ-த்த்-உ.வே.பி.

மரம்+ஐ > மரத்தை

மரம்+ஆல் > மரத்தால்

மரம்+உக்கு > மரத்துக்கு

மரம்+ஓடு > பொன்னோடு

மரம்+உடன் > மரத்துடன்

மரம்+இன் > மரத்தின்

மரம்+உடைய > மரத்துடைய

மரம்+அது > மரத்தினது

மரம்+இல் > மரத்தில்

6. *அ* ஈற்றுப் பெயர்களான *பல, சில* என்பனவற்றைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டுகள் வரும்போது *அற்ற்* என்ற சாரியை இடையில் வரும்.

பல + உ.வே.பி.(பல-வ்-அற்ற்-உ.வே.பி.

பல+ ஐ > பலவற்றை

பல+ஆல் > பலவற்றால்

பல+உக்கு > பலவற்றுக்கு

பல+ஓடு>பலவற்றோடு

பல+உடன் > பலவற்றுடன்

பல+இன் > பலவற்றின்

பல + உடைய > பலவற்றுடைய

பல + அது > பலவற்றினது

பல + இல் >பலவற்றில்

8. எல்லாம் என்ற பெயரைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டுக்கள் வரும்போது *ம்* கெட *அற்ற்* என்ற சாரியை இடையில் வரும்.

எல்லாம் + உ.வே.பி.→எல்லா-வ்-அற்ற் – உ.வே.பி.

எல்லாம் + ஐ > எல்லாவற்றை

எல்லாம் +ஆல் > எல்லாவற்றால்

எல்லாம் +உக்கு > எல்லாவற்றுக்கு

எல்லாம் +ஓடு>எல்லாவற்றோடு

எல்லாம் +உடன் > எல்லாவற்றுடன்

எல்லாம் +இன் > எல்லாவற்றின்

எல்லாம் +உடைய > எல்லாவற்றுடைய

எல்லாம் +அது > எல்லாவற்றினது

9. 5.ஆம் 6- ஆம் விதிகளுக்கு உட்படாத *ண், ல், ள், ன்* மற்றும் *ர், ழ்* என முடியும் பெயர்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டுக்கள் வரும்போது அவை எந்தவித மாற்றமுமின்றி இணையும் (விதிவிலக்கு: *ய் + உக்கு > ய்க்கு*).

(5-ஆம், 6-ஆம் விதிக்கு உட்படாத) *-ண்/ ல்/ ள்/ ன்/ர்/ ழ்* + உ.வே.பி. (ண்/ ல்/ ள்/ ன்/ர்/ ழ் + உ.வே.பி.

எ.கா

தூண்+ ஐ > தூணை

தூண்+ஆல் >துணால்

தூண்+உக்கு > தூணுக்கு

தூண் +ஓடு >தூணோடு

தூண் + உடன் > தூணுடன்

தூண்+இன் > தூணின்

தூண்+உடைய >தூணுடைய

தூண்+அது > தூணது

தூண்+இல் > தூணில்

நாய்+ ஐ > நாயை

நாய்+ஆல்>நாயால்

நாய்+உக்கு>நாய்க்கு

நாய்+ஓடு>நாயோடு

நாய்+உடன் > நாயுடன்

நாய்+இன் > நாயின்

நாய் + உடைய > நாயுடைய

நாய் + அது > நாயது

நாய் + இல் > நாயில்

கால் + ஐ > காலை

கால் + ஆல் > காலால்

கால் + உக்கு > காலுக்கு

கால் + ஓடு > காலோடு

கால் + உடன் > காலுடன்

கால் + இன் > காலின்

கால்+உடைய > காலுடைய

கால்+அது > காலது

கால்+இல் > காலில்.

வாள்+ ஐ > வாளை

வாள்+ஆல்>வாளால்

வாள்+ உக்கு > வாளுக்கு

வாள்+ ஓடு >வாளோடு

வாள்+ உடன் > வாளுடன்

வாள்+ இன் >வாளின்

வாள்+ உடைய > வாளுடைய

வாள்+ அது>வாளது

வாள்+ இல்>வாளில்

மான்+ ஐ > மானை

மான்+ ஆல்>மானால்

மான்+ உக்கு >மானுக்கு

மான்+ ஓடு>மானோடு

மான்+ உடன் >மானுடன்

மான்+ இன்>மானின்

மான்+ உடைய >மானுடைய

மான்+ அது>மானது

மான்+ இல் > மானில்.

சுவர்+ ஐ > சுவரை

சுவர்+ ஆல் > சுவரால்

சுவர்+உக்கு > சுவருக்கு

சுவர்+ஓடு > சுவரோடு

சுவர்+ உடன் > சுவருடன்

சுவர்+ இன் > சுவரின்

சுவர்+ உடைய > சுவருடைய

சுவர்+ அது > சுவரது

சுவர்+இல்>சுவரில்.

இதழ்+ ஐ > இதழை

இதழ்+ ஆல்>இதழால்

இதழ்+ உக்கு >இதழுக்கு

இதழ்+ ஓடு >இதழோடு

இதழ்+ உடன் >இதழுடன்

இதழ்+ இன் >இதழின்

இதழ்+ உடைய >இதழுடைய

இதழ்+ அது>இதழது

இதழ்+ இல்>இதழில்.

10. *இன்* என்ற சாரியை ஒருமைப்பன்மைப் பெயர்ப்பகுதிகளைத் தொடர்ந்து வேற்றுமை உருபுகளுக்கு முன்னர் விருப்பாக வரும்.

எ.கா.

பசு	+	ஐ	>பசுவினை
பசு	+	ஆல்	>பசுவினால்
பசு	+	உக்கு	>பசுவினுக்கு
பசு	+	ஓடு	>பசுவினோடு
பசு	+	உடன்	>பசுவினுடன்
பசு	+	உடைய	>பசுவினுடைய
பசு	+	அது	>பசுவினது
பசு	+	இல்	>பசுவினில்.

மேற்கண்ட உருபொலியனியல் விதிகள் அடிப்படையில் தான் இராசேந்திரன் (Rajendran 1999) தன் உருபனியல் பகுப்பாய்வியின் உருவாக்கத்திற்காகப் பெயர்களை 23 வகைகளாகப் பிரித்துள்ளார். இவ்வாறு சொல் அடுக்குகளைப் பெயர் அடுக்குகள் (nouns pardigms) வினையடுக்குகள் (verb paradigms), பெயரடை அடுக்குகள் (adjective pardigms) வினையடை அடுக்குகள் (verb paradigms), எண்ணடுக்குகள் (numeral paradigms) வகைப்படுத்தி உருபனியல் மாற்றங்களைக் கணினிக்கு அறியச்செய்து சொற்களை உருபன்களாகப் பிரிக்கும் அணுகுமுறை சொல்லடுக்கு அணுகுமுறை (paradigm approach) அடிப்படையிலானதாகும். இத்தகைய உருபனியல் பகுப்பாய்விதான் இந்திய அரசின் நிதி நல்கையில் நடைபெற்ற இந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயர்ப்புக்கான உருபனியல் பகுப்பாய்வியில் பயன்பத்தப்பட்டுள்ளது. இத்தகைய ஆய்வி குறித்து (Rajendran 1999) தனது கட்டுரையிலும்

வைஷணவி (2003) தனது ஆய்வேட்டில் விரிவாகத் தந்துள்ளனர். இத்தகைய அணுகுமுறைதான் தெய்வசுந்தரமும் சந்தி பிரிப்பான், உருபனியல் ஆய்வி, பிழைதிருத்தி ஆகிய கருவிகள் கொண்ட 'மென்தமிழ்" என்று அழைக்கப்படுகின்ற தமது சொல்லாளர் (word processor) உருவாக்கத்திற்குப் பயன்படுத்துகின்றார்.

4.3.2.7. வினை அடுக்குகள்

வினைகளை அவை ஏற்கும் இறந்தகால–நிகழ்கால–எதிர்கால ஒட்டுகளின் அடிப்படையில் ஏழு பெரும் பிரிவுகளாகவும் உருபனியல் விதிகளின் அடிப்படையில் மேலும் சிறு பிரிவுகளாகப் பிரித்து சீரற்றுத் திரிபுறும் வினைகளையும் கருத்திற்கொண்டு 27 பிரிவுகளாகப் பிரிக்கலாம். அவையாவன: செய் - பிரிவு, அழு – பிரிவு, சாகு – பிரிவு, விடு – பிரிவு, பெறு – பிரிவு, வே –பிரிவு, புகழ்; - பிரிவு, நோ – பிரிவு, அறி – பிரிவு, அகல் - பிரிவு, கொல் - பிரிவு, ஆள் - பிரிவு, கொள் - பிரிவு, ஓடு – பிரிவு, சொல் - பிரிவு, போ – பிரிவு, கல் - பிரிவு, கேள் - பிரிவு, உண் - பிரிவு, பூண் - பிரிவு, காண் - பிரிவு, தின் - பிரிவு, நில் - பிரிவு, படி - பிரிவு, நட – பிரிவு. இத்தகைய வினை அடுக்குகள் தான் சொல் அடுக்குமுறை உருபனியல் கோட்பாடு அடிப்படையிலான அடுக்கு அணுமுறைக்கு அடிப்படையாக அமைகின்றது. முன்னர் கூறியப்டி வினைகளை வினையடுக்குகளாகக் குழுமி கணிப்பொறிக்கு உருபொலியனியல் விதிகளை கற்பித்து உருபனியல் ஆய்வு செய்யும் உருபனியல் ஆய்வி உருவாக்கப்படுகின்றது.

த்/கிற்/வ் பிரிவு

இதில் அடங்கும் வினைகளை ஐந்து உட்பிரிவுகளாகப் பாகுபாடு செய்யலாம்.

வ.எண்	வினைபகுதி	இறந்தகால வி.மு. ஆண்பால்	நிகழ்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஒன்றன்பால்
1	செய்	செய்தான்	செய்கிறான்	செய்வான்	செய்யும்
2	அழு	அழுதான்	அழுகிறான்	அழுவான்	அழும்
3	சாகு	செத்தான்	சாகிறான்	சாவான்	சாகும்
4	விடு	விட்டான்	விடுகிறான்	விடுவான்	விடும்
5	பெறு	பெற்றான்	பெறுகிறான்	பெறுவான்	பெறும்

ந்த்/கிற்/வ் பிரிவு

இதில் 10 உட்பிரிவுகள் வரும்.

வ.எண்	வினைபகுதி	இறந்தகால வி.மு. ஆண்பால்	நிகழ்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஒன்றன்பால்

1.	புகழ்	புகழ்ந்தான்	புகழ்கிறான்	புகழ்வான்	புகழும்
2.	விழு	விழுந்தான்	விழுகிறான்	விழுவான்	விழும்
3.	அறி	அறிந்தான்	அறிகிறான்	அறிவான்	அறியும்
4.	நோ	நொந்தான்	நோகிறான்	நோவான்	நோகும்
5.	வே	வெந்தான்	வேகிறான்	வேவான்	வேகும்
6.	வா	வந்தான்	வருகிறான்	வருவான்	வரும்
7.	அகல்	அகன்றான்	அகல்கிறான்	அகல்வான்	அகலும்
8.	கொல்	கொன்றான்	கொல்கிறான்	கொல்வான்	கொல்லும்
9.	ஆள்	ஆண்டான்	ஆள்கிறான்	ஆள்வான்	ஆளும்
10.	கொள்	கொண்டான்	கொள்கிறான்	கொள்வான்	கொள்ளும்

த்/கிற்/ப்/ பிரிவு

வ.எண்	வினைபகுதி	இறந்தகால வி.மு. ஆண்பால்	நிகழ்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஒன்றன்பால்
1.	ஓடு	ஓடினான்	ஓடுகிறான்	ஓடுவான்	ஓடும்
2.	சொல்	சொன்னான்	சொல்கிறான்	சொல்வான்	சொல்லும்
3.	போ	போனான்	போகிறான்	போவான்	போகும்

த்/கிற்/ப்/பிரிவு

வ.எண்	வினைபகுதி	இறந்தகால வி.மு. ஆண்பால்	நிகழ்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஒன்றன்பால்
1.	கல்	கற்றான்	கற்கிறான்	கற்பான்	கற்கும்
2.	கேள்	கேட்பான்	கேட்கிறான்	கேட்பான்	கேட்கும்
3.	உண்	உண்டான்	உண்கிறான்	உண்பான்	உண்ணும்
4.	பூண்	பூண்டான்	பூணுகிறான்	பூண்பான்	பூணும்
5.	காண்	கண்டான்	காண்கிறான்	காண்பான்	காணும்
6.	தின்	தின்பான்	தின்கிறான்	தின்பான்	தின்னும்

ந்த்/கிற்/ப்/ பிரிவு

வ.எண்	வினைபகுதி	இறந்தகால வி.மு. ஆண்பால்	நிகழ்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஒன்றன்பால்

1.	நில்	நின்றான்	நிற்கிறான்	நிற்பான்	நிற்கும்

த்த்/க்கிற்/ப்ப்/ பிரிவு

வ.எண்	வினைபகுதி	இறந்தகால வி.மு. ஆண்பால்	நிகழ்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஒன்றன்பால்
1.	படி	படித்தான்	படிக்கிறான்	படிப்பான்	படிக்கும்

ந்த்/ க்கிற்/ ப்ப்/ பிரிவு

வ.எண்	வினைபகுதி	இறந்தகால வி.மு. ஆண்பால்	நிகழ்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஆண்பால்	எதிர்கால வி.மு. ஒன்றன்பால்
1.	நட	நடந்தான்	நடக்கிறான்	நடப்பான்	நடக்கும்

இராசேந்திரன் உருவாக்கிய உருபனியல் ஆய்வி அதன் உருவாக்கதிற்காக பெயர்களை 23 வகைகளாகப் பிரிக்கிறார். இப்பிரிவுகளின் அடிபடையில் தான் அவருடைய உருபனியல் உருவாக்கிச் செயல்படுகின்றது.

4.4 தமிழில் உருபனியல் ஆய்விகள்

தமிழ்மொழி ஒரு திராவிட மொழி. இம்மொழி வினைச்சொல்லை இறுதியாகக் கொண்ட வாக்கியங்களை உடையதாகும். இம்மொழி இலக்கணச் செயற்பாங்குகளை (அதாவது இலக்கணப் பொருண்மைகளை) உருபனியலில் நிலையில் வெளப்படுத்தும் ஒட்டுக்கள் நிறைந்த மொழியாக விளங்குகிறது. இம்மொழியில் உள்ள ஒவ்வொரு வேர் உருபனும் ஆயிரக்கணக்கான சொற்களின் உருவாக்கத்திற்கு (அதாவது திரிபு மற்றும் ஆக்க வடிவுகளின் உருவாக்கத்திற்கு) அடிப்படையாக அமைகிறது. தமிழ் வாக்கியங்கள் எழுவாய்தொடரையும் வினையையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் செயப்படுபொருள் தொடரையும் வினையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் பெயர்த்தொடரையும் (அதாவது பயனிலைத் தொடரையும்) கொண்ட அமைப்பாகவோ வரும். இதன் காரணமாகத் தமிழ் மொழி ஒரு SOV மொழி என்பர். தமிழ் மொழி ஒரு முன் மாதிரி SOV மொழிக்குள்ள எல்லாப் பண்புக்கூறுகளையும் பெற்றுள்ளது. பெயர்கள் வேற்றுமை உருபுகளுக்காகத் திரிபுறுகின்றன. பெயர்த்தொடர்களுக்கும் வினைகளுக்கும் உள்ள உறவுகள் வேற்றுமை உருபுகளாலோ பின்னுருபுகளாலோ வெளிப்படுத்தப்படுகின்றன. வினைகள் கால இடைநிலைகளை ஏற்பதாகவும் எழுவாய்த்தொடருக்கு உடன்பாடாக வினைமுற்று விகுதிகளை ஏற்பதாகவும் வருகின்றன. இதன் காரணமாக எந்தக் கணிப்பொறி தொடர்பான மொழி ஆய்விற்கும் உருபனியல் பகுப்பாய்வு இன்றியமையாததாகவும் அடிப்படையாகவும் அமைகின்றது. மேலும் எழுத்துப்பிழை திருத்தி உருவாக்கத்திற்கும் இலக்கணப்பிழை திருத்தி உருவாக்கத்திற்கும் உருபனியல் பகுப்பாய்வு அடிப்படையாகும்.

உருபனியல் ஆய்வியின் உருவாக்கம் எந்த இயற்கை மொழி ஆய்வின் செயல்பாட்டிற்கும் மிக முக்கியமான முதற்படியாகும். வேறுபட்ட மொழி வகைகளுக்கு வேறுபட்ட ஆய்வு

நெறிமுறைகள் இருக்கின்றன. எளிய இணைப்பு உருபனியல் உள்ள (simple concatinative morphology) மொழிகள் சொல்லும் செயற்பாங்கும் (item and process) மற்றும் சொல்லும் வரிசைமுறையும் (item and arrangement) என்ற மொழி மாதிரிகளால் உருவாக்கப்பட்ட ஆய்விகளால் தமது தேவையைப் பூர்த்தி செய்துகொண்டுள்ளன. பண்புக்கூறுகள் அடிப்படையிலான ஒருமைப்படுத்தும் கோட்பாடுகள் (unification theories) ஸ்பானிஷ் போன்ற மொழிகளுக்குப் பயன்படுத்தப்பட்டது. இக்காலத்தைய மிகப் புகழ்வாய்ந்த கணினிசார் இலக்கணமான தலைமையால் இயக்கப்படும் தொடரமைப்பு இலக்கணம் (Head-Driven Phrase Structure Grammar (HDPSG) பண்புக்கூறு அடிப்படையிலான உருபனியலைப் (Feature Basecd Morphology) பயன்படுத்துகின்றது. இங்கு தமிழ்மொழியின் உருபனியல் ஆய்வுக்குப் பயன்படுத்தப்பட்ட மாதிரிகளும் வேறு முன்னணி மாதிரிகளும் விளக்கப்படும்.

கணிப்பொறி வழி ஆய்வைப் பொறுத்தவரையில் இருமட்ட மாதிரி (two level model) உருபனியலின் முக்கிய கோட்பாடுகளில் ஒன்றாகும். இது பரந்த பரப்பெல்லையில்படும் மொழிகளுக்குச் செயல்படுத்தப்பட்டிருக்கின்றது. முற்றுநிலை தானியங்கி (Finite State Automata (FSA)) மாதிரி சென்னையிலுள்ள AUKBC ஆய்வு மையத்தால் உருவாக்கப்பட்ட உருபனியல் பகுப்பாய்வி முற்றுநிலைத் தானியங்கி அட்டவணையைப் (finite automata state table) பயன்படுத்தி செயல்படுகின்றது. தலைமைத் தொடரில் அமைப்பு இலக்கணம் (Head Phrase Structure Grammar (HPSC)) தமிழுக்கு உமாராணியால் (Umarani, 2001) பயன்படுத்தப்பட்டுள்ளது. ஆம்பிள் என்ற மற்றொரு மாதிரி சொல்லும் வரிசையும் என்ற மாதிரியுடன் தொடர்புடையது. பகுத்துலும் உருவாக்கலும் (Parsing and Generation) இயற்கை மொழிகளின் உருபனியல் ஆய்வின் இரு முக்கியமான நேர்வுகள் ஆகும். பகுப்பாய்வு என்பது பகுப்பாய்வி தரப்பட்ட உள்ளீட்டை புரிந்து கொள்ளவும் ஆயவும் செய்து பெரும்பாலும் அவற்றைப் பொருண்மை மற்றும் இலக்கணப் பண்புக் கூறுகளுடன் வெளியீடாகத் திருப்பித்தரும் செயல்பாடாகும். பல உரை மீட்பு ஒழுங்குமுறைகளுக்கும் (Text Retrieval Systems (TRS)) இயந்திர மொழிபெயர்ப்புச் (Machine Translation (MT)) செயல்பாடுகளுக்கும் பகுப்பாய்வு முக்கியமானது ஆகும். அதாவது இந்த இரண்டு செயல்பாடுகளுக்கும் தொடக்க இடம் பகுப்பாய்வாகும். ஒரு உணரி (recognizer) உள்ளீட்டுச் சொல்லை வேறுபட்ட உருபன்களாகப் பகுப்பாய்வு செய்கின்றது. ஆங்கிலச் சொல்லான *reads* என்பது பின்வருமாறு புரிந்து கொள்ளப்படும்: *read* என்ற வினையில் படர்க்கை இட ஒருமை எண் குறிப்பிடப்பட்ட நிகழ்கால வடிவம். பகுப்பாய்வி உணரியிலிருந்து அது புரிந்து கொள்ளும் சொற்களுக்குப் பொருத்தமான அமைப்பைத் தருகிறது என்ற அளவில் வேறுபடுகின்றது. இவ்வாறு ஒரு உருபனியல் பகுப்பாய்வி உள்ளீடு செய்யப்பட்ட சொல்லை அதன் உறுப்பு உருபன்களாகப் பிரித்து அதனைப் படிநிலை அமைப்பாகப் பகுப்பாய்வு செய்கிறது. இது சொல்லாக்கத்தில் உருபன்களை இணைப்பதற்கு மொழியில் குறிப்பிட்ட இயக்க

நெறியின் ஒரு நோக்கை வெளிப்படுத்துகிறது. இவ்விக்கநெறி ஒரு வேருடன் கட்டுண்ட உருபுகளை முன்னொட்டாகவோ உள்ளொட்டாகவோ பின்னொட்டாகவோ வேருடன் சேர்க்கும். தமிழ்மொழியில் இது கூடுதலாகப் பின்னொட்டாக்கம் ஆகும். முன்னொட்டுக்கள் சமஸ்கிருதத்தில் இருந்து வந்த சில கடன் சொற்களில் இருக்கின்றன. மாறாக, உருபனியல் உருவாக்கி (Morphological Generator) உருபன்களை இணைத்துச் சொற்களை உருவாக்கும் நோக்கைப் பயன்படுத்துகிறது. இவ்வாறு உருபனியல் உருவாக்கி பகுப்பாய்வதன் மறுதலையாகும். இதன் உள்ளீடு ஒரு குழும உருபன்களோ அதன் இலக்கண அமைப்புகளோ ஆகும். இது அவற்றை ஒன்றிணைத்து ஒரு மொழியில் நல்வடிவாக்கமுள்ள சொற்களை உருவாக்குகிறது. உருபனியல் பகுப்பாய்வி இச்செயல்பாடுகளில் ஒன்றையோ அதற்கு மேலானவற்றையோ செய்கின்றது. வேறுபட்ட உருபனியல் ஆய்விகள் அவை விளக்கும் மொழிகளின் அடிப்படையில் வேறுபட்ட கோட்பாடு அணுகுமுறைகளைப் பயன்படுத்துகிறது. அதாவது வேறுபட்ட மொழிகளின் இயல்பும் கலவைத்தன்மையும் அவற்றை ஆயப்பொருத்தமான கோட்பாடு அணுகுமுறைகளைத் தீர்மானிக்கிறது. வேறுபட்ட சட்டங்களின் அடிப்படையில் வேறுபட்ட உருபனியல் ஆய்விகள் உருவாக்கப்படுகின்றன. வேறுபட்ட சட்டகங்களின் தோ;வும் அவற்றின் வேறுபட்ட நடைமுறைபடுத்தல்களும் ஒப்பீடு செய்யும் ஆய்வுக்கு அவசியமாகிறது.

இப்பகுதியில் இருமட்ட உருபனியலைப் பயன்படுத்தும் பிசிகிம்மோ (PCKimmo), ஆம்பிள் (Ample) என்பனவும் சொல்லும் வரிசைமுறையும் என்ற மாதிரியின் அடிப்படையில் அமைந்த ஜிஎஸ்மார்ஃபும் (GSMorph) சொல்லும் செயற்பாங்கும் மாதிரியின் அடிப்படையில் அமைந்த இராமசாமியின் (Ramaswamy, 2000) தமிழ் உருவாக்கியும் (Generator for Tamil) சொல் ஒலியனியல் (Lexical Phonology) அணுகுமுறை அடிப்படையில் அமைந்த இரங்கநாதனின் (Renganathan,1997) டேக் தமிழும் (Tag Tamil) சொல்லும் வரிசைமுறையும் மற்றம் சொல்லும் செயற்பாங்கும் என்ற இரு மாதிரிகளையும் பயன்படுத்தும் வைஷ்ணவியின் தமிழ்; உருபனியல் பகுப்பாய்வியும் உருவாக்கியும் இராசேந்திரனின் உருபனியல் ஆய்வியும் விளக்கப்படுகின்றன. இவை தவிர தமிழில் முயற்சிக்கப்பட்டுள்ள பிற உருபனியல் ஆய்விகளும் பட்டியலிடப்பட்டுள்ளன.

தமிழில் கணிப்பொறி தொடர்பான செயல்பாடுகளுக்கு அடிப்படையாக அமைவது உருபனியல் பகுப்பாய்வும் உருவாக்கமும் ஆகும். பலர் இதில் பங்களிப்பு செய்துள்ளனர். பல உருபனியல் ஆய்விகள் தமிழில் உருவாக்கப்பட்டுள்ளன. இவற்றில் முக்கியமானவைகளைப்பற்றிய சிறு குறிப்பு கீழே தரப்பட்டுள்ளன.

4.4.1. இராஜேந்திரனின் தமிழ் உருபனியல் பகுப்பாய்வி

முதல் முதலில் இந்திய மொழிகளுக்கிடையில் இயந்திர மொழி பெயர்ப்பை மேற்கொள்ள வேண்டி அனுசாராகா (anusaraka 'language acessor') என்ற மொழிபெயர்புக் கருவி உருவாக்கத்திற்காகப் பாணினி வடிவவாதம் அடிப்படையில் ஆய்வு மேற்கொண்ட குழுவைச் சார்ந்த ஆய்வாளர்களால் DOE நிதி நல்கையில் தொடங்கிய ஆய்வுதான் இந்தியாவில் இயற்கை

மொழி ஆய்விற்கு வித்திட்டது எனலாம். இந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயர்ப்பு ஆய்வுக்காக உருவாக்கப்பட்ட உருபனியல் பகுப்பாய்விகள் அடிப்படையில் தமிழ்-இந்தி மொழி பெயர்ப்புக்காகத் தமிழ் உருபனியல் பகுப்பாய்வியும் இராசேந்திரன் உதவியுன் உருவாக்கப்பட்டது. AUKBC ஆய்வு நிறுவனக் குழுவினரால் உருவாக்கப்பட்ட அபி உருபனியல் பகுப்பாய்வியும் இராசேந்திரனின் மேற்பார்வையில் உருவாக்கப்பட்டதுதான். தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையில் பல்கலைக்கழக மானியக் குழுவின் நிதிநல்கையில் கீழ் மேற்கொள்ளப்பட்ட "Spell and Grammar Checher for Tamil" என்ற உயராய்வுக்காக PROLOG வழியமைப்பு மொழியைப் பயன்படுத்தி ஒரு உருபனியல் பகுப்பாய்வி இராசேந்திரனால் உருவாக்கப்பட்டது. இப்பகுப்பாய்வி இந்திய மொழிகளின் நடுவண் நிறுவன விரிதரவில் பயன்படுத்தப்பட்டு மதிப்பீடு செய்தபோது 95% வெற்றிகரமாக பகுத்துக்குறித்தல் வெளியீடு கிடைத்தது.

முன்னர் குறிப்பிட்டது போன்று இந்த உருபனியல் பகுப்பாய்வி சொல்லடுக்கு அணுகுமுறையை (word and paradigm) அடித்தளமாகக் கொண்ட அடுக்கு அணுகுமுறை (paradigm approach) அடிப்படையில் உருவாக்கப்பட்டுள்ளது.

4.4.2. கணேசனின் உருபனியல் பகுப்பாய்வி

கணேசன் (Ganesan, 1994, 2003) அவர்கள் தமிழுக்காக இந்திய மொழிகளின் மையநிறுவனம் உருவாக்கிய மூன்று மில்லியன் சொற்கள் அடங்கிய விரிதரவை (CIIL Corpus) ஆய வேண்டி ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இவர் ஒலியனியல் மற்றும் உருபனியல் விதிகளைக் கையாண்டு இப்பகுப்பாய்வியை உருவாக்கினார். சமீபத்தில் சிறந்த தமிழ் உருபனியல் பகுப்பாய்வியை உருவாக்கியுள்ளார்.

4.4.3. கபிலனின் தமிழ் வினைகளுக்கான உருபனியல் பகுப்பாய்வி

கபிலன் (Kabilan, 1994) அவர்கள் தமிழ் வினைகளுக்கான ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இது பேராசிரியர் தெய்வசுந்தரத்தின் (மொழியியல் பகுதி, தமிழ்மொழித் துறை, சென்னைப் பல்கலைக்கழகம்) மேற்பார்வையின் கீழ் நடந்த தமிழ்சார் இயற்கை மொழி ஆய்வின் விளைவாகும்.

4.4.4. தெய்வசுந்தரத்தின் உருபனியல் பகுத்துக் குறிப்பான்

தெய்வசுந்தரம் தமிழுக்கென ஒரு சொல்லாய்வியைத் (Word Processor) தயாரித்துள்ளார். இது மிக நன்றாகச் செயல்பாடு செய்கின்றது. இது இருநிலை மாதிரியை (two-level-model) அடிப்படையாகக் கொண்டது. புணர்ச்சி விதிகள் பகுப்பாய்வுக்குப் பயன்படுத்தப்படுகின்றன. இவர் ஆய்வி (Deivasundaram and Gopal 2003) உருபொலியனியல் மற்றும் உருபனியல் என்ற இருநிலைகளில் செயல்படுகின்றது. இது ஒலியன் சேர்ப்பையும் (phonemic addition), ஒலியன் மாற்றத்தையும் (phonemic change) கணக்கிலெடுக்கின்றது. இவ்வாய்வி உருபனியல்

அமைப்பொழுங்கு (mophotactics) அடிப்படையில் ஒட்டுகள் பகுதியுடன் இணைக்கப்படும் நிரலைக் கணக்கிலெடுத்துக் கொண்டு பகுப்பாய்வு செய்யும். இது ஒரு எடுத்துக்காட்டான உருபனியல் பகுப்பாய்வியாகும். முற்றிலும் மொழியல் விதிகளைப் பயன்படுத்தி வெற்றிகரமாக முடிக்கப்பட்டு தமிழுக்கான தலைச் சிறந்த சொல்லாய்வியாக (சொல்லாளரகாக/மென் தமிழாக) வளர்து இது தமிழ் உலகை வலம் வருகின்றது. அரசுப்பணிகளுக்காக இச்சொலாய்வி தமிழக அரசால் பரிந்துரைக்கப்பட்டுப் பரவலாகப் பயன்படுத்தப்பட்டு வருகின்றது.

4.4.5. AUKBC நிறுவனத் தமிழ் உருபனியல் பகுப்பாய்வி

AUKBC ஆய்வு நிறுவனம் (தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையுடன் இணைந்து) உருவாக்கிய API உருபனியல் சொல்லாய்வி தரவுக்கோப்பு, நிலை மாற்றமைவு அட்டவணை (State table.dat) என்பனவற்றில் தந்துள்ள தானியங்கி அட்டவணையைச் சார்ந்து இருக்கிறது. API உருபனியல் சொல்லாய்வி சொற்களை அதன் திரிபுகளிலிருந்து பிரித்தெடுக்கும்படி வடிவமைக்கப்பட்டுள்ளது.

4.4.6. வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும்

வைஷ்ணவி (Vaishnavi Ramaswamy, 2003) முனைவர் பட்ட ஆய்வுக்காகத் தமிழ் உருபனியல் பகுப்பாய்வை எடுத்துக்கொண்டு ஆய்வேட்டினைச் செய்துள்ளார். மேலும் வைஷ்ணவி (Vaishnavi Ramaswamy, 2002) தமது ஆய்வியல் நிறைஞர் ஆய்வேட்டிற்காக தமிழுக்கு ஒரு உருபனியல் உருவாக்கியை உருவாக்கி விளக்கமாகத் தந்துள்ளார். அவருடைய உருபனியல் உருவாக்கி சொல்லும் வரிசைமுறையும் (word and arrangement model) மாதிரியையும் சொல்லும் செயற்பாங்கும் மாதிரியையும் (word and process model) சொல்லும் அடுக்கும் (word and paradigm model) ஆகிய மூன்று கோட்பாடுகளையும் பயன்படுத்தி மேற்கொள்ளப்பட்ட ஒரு கலப்பு அணுகுமுறைகாகும். இந்த உருவாக்கி பிசிகிமோவின் (Pckimmo) கூட்டிணைப்பு நெறிமுறையால் செயல்படுகின்றது. வைஷ்ணவி இதைக் கலப்பு மாதிரி (hybrid model) என்று அழைக்கின்றார்.

PERL–இல் உருவாக்கப்பட்ட மற்றும் செயல்படுத்தப்பட்ட கலப்பு மாதிரி (hybrid model) உள்ளுருவாக்கப்பட்ட சொற்களஞ்சியத்தின் சட்டகத்தையும் மொழியின் வேறுபட்ட ஒவ்வொரு சொல் வகுப்பிற்கும் சொல்லும் வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் என்பனவற்றின் ஒரு குழும விதிகளையும் கொண்டிருக்கிறது. தரப்பட்ட உருபொலியனியல் விதிகளின் அடிப்படையில் செயல்படும் இணையான துணை நடைமுறைகளை (corresponding subroutines) வேண்டுவதன் மூலம் இதன் ஆய்வு நிரூபிக்கப்படுகிறது. இது ஒவ்வொரு வகைப்பாட்டின் சொல் வடிவுகளைக் கையாளும்படி வடிவமைக்கப் பட்டுள்ளது. இது முதன்மை அமைப்பின் கோப்புகளால் கட்டளையிட்ட படி செயல்படுகிறது. வழியமைப்பு மொழியில் மொழித் தகவலின் மரபுரிமை குறியாக்கம் இருக்கிறது. திறந்த இறுதியுள்ள உருபொலியனியல்

மற்றும் மாற்றுருபு விதிகளின் (open ended morphophonimic and allomorphic rules) வழி மொழித் தகவலின் பகுப்பாய்வால் இது நடைப்பெறுகிறது (Vaishnavi Ramaswamy, 2003:101-110).

இம்மாதிரி உருபனியல் சொல்லும்வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் மாதிரிகளின் கலவையில் அதன் கோட்பாட்டு அடிப்படையை அமைத்துக் கொள்கிறது. செயற்பாட்டைச் சொற்பகுதியில் (word stem) உருபனியல் இயல்பைப் பொறுத்து ஒரு மாற்றத்தை (transformation) விவரணை செய்வதாக விளக்க இயலும். திரிபு நீக்கத்தின் (inflectional stripping) ஒவ்வவொரு நிலையிலும் குறிப்பிடப்பட்டுள்ள கட்டுப்பாடுகளின் முழுத்தொடர்சிகளின் வழி கடந்த பின்னர் ஒரு சொல் பகுப்பாய்வு செய்யப்படுகிறது. ஒரு பகுப்பாய்வுத் தொகுதி (analysis module) வலமிருந்து இடமாகக் கோர்வை அமைவதை வருடிச் (scanning) செய்கிறது. ஒரு நேரத்தில் ஒரு பின்னொட்டை வருடி அறிந்து பிரித்தெடுக்கிறது. மற்றும் ஒவ்வொரு நேர்விலும் எடுத்துக்காட்டப்பட்ட ஒலியனியல் மற்றும் உருபொலியனியல் விதிகளின் உதவியுடன் மீதிச்சொல்லை மீட்டுருவாக்கம் செய்கிறது. கோர்வை தீர்வது வரை இது நடைபெறும். வெளியீடு கண்டுபிடிக்கப்பட்ட உருபன்களின் வடிவில் இருக்கும். பின்னர் அவை உறுதி செய்யப்படும்.

இவர் தமது உருபனியல் ஆய்வி உருவாக்கத்திற்காக உருபொலியனியல் விதிகள் அடிப்படையில் பெயர்களை 24 குழுக்களாகவும் மாற்றுப்பெயர்களை 15 குழுக்களாகவும் வினைகளை 24 குழுக்களாகவும் பெயரடைகளை இரண்டு குழுக்களாகவும் பகுப்பாய்வு செய்து கணினிக்கு உருபொலியனியல் விதிகளை அறியச்செய்து தமிழ் உருபனியல் பகுப்பாய்வியை உருவக்கியுள்ளார்.

4.4.7. வின்ஸ்டன் குருசின் ஜி.எஸ். மார்ஃப் பகுப்பாய்வியும் உருவாக்கியும்

ஜி.எஸ். மார்ஃப் (GS Morph) என்று சுருக்கமாகக் குறிப்பிடபப்படும் எழுத்து அடிப்படையிலான அமைப்பு மொழியியலார் உருபனியல் பகுப்பாய்வி (Grapheme Oriented Structuralist Morphological Processor (GS Morph)) (Sengupta 1999) ஆம்பிள் போன்று சொல்லும் வரிசை முறையும் மாதிரியின் அடிப்படையில் அமைந்தது. இது ஒவ்வொரு உருபனின் விரிவான புற மெய்ப்படுத்தத்தை வேண்டும் அல்லது ஏற்படுத்தும். ஆம்பிள் ஒட்டுக்களை அவை பின்னொட்டுக்களா உள்ளொட்டுக்களா என்பதன் அடிப்படையில் வேறுப்பட்ட கோப்புகளில் சேகரிக்க இயலும். ஆனால் ஜிஎஸ் மார்ஃபின் பகுப்பாய்வு ஒரு வேர்ச் சொற்களஞ்சியத்தையும் (Root Lexicon) பின்னொட்டுக்களின் சொற்களஞ்சியத்தையும் (Suffix Lexicon) உட்படுத்தும்.

ஜிஎஸ் மார்ஃப் சொல்லடைவைப் பொருத்துவதால் செயல்படுகிறது. அதன் உள்ளீடு எழுத்து உருப்படுத்தத்தின் வடிவில் இருக்கிறது. அது வலது எல்லையிலிருந்து இடது எல்லைக்குப் பகுப்பாய்வைத் தொடங்குகிறது. முதலாவது அது தரப்பட்ட உள்ளீட்டைப் பூஜிய பின்னொட்டால் தொடரப்படும் வேராகக் கருதுகிறது. இதன்படி இது பின்னொட்டுச் சொற்களஞ்சியத்தில் பின்னொட்டுக்காகப் பரிசோதிக்கும். பின்னர் அது வோ;க்கோப்பில் பூஜிய ஒட்டிற்கு நேராக இருக்கும் சொல்லடைவுகளைப் (concordance) பார்க்கும். அது உள்ளீட்டைப்

பொருத்துவதில் தோல்வியுற்றால் அது சொல் வடிவத்தைப் பற்றிய அதன் ஆரம்ப அனுமானத்தை மீளப்பெறும். அது உள்ளீட்டின் எழுத்துக்களை வெட்டி எடுக்கத் தொடங்கும். ஒவ்வொரு தடவையும் ஒரு எழுத்தை வெட்டி எடுக்கும்போது அது எல்லா வெட்டி எடுக்கப்பட்ட எழுத்துக்களும் சேர்ந்த ஒன்றைப் பின்னொட்டுச் சொற்களஞ்சித்தில் பொருத்தத்திற்காகப் பார்க்கும். வெட்டி எடுக்கும் செயற்பாங்கு முழுகோர்வையும் தீருவது வரை தொடர்ந்து நடைபெறும். பின்னொட்டுச் சொற்களஞ்சியத்தில் தேடல் வெற்றி பெற்றால் உள்ளீட்டின் மீதிப்பகுதி வேராக முன்மொழியப்பட்டு வேர்க்கோப்பில் பார்க்கப்படும். முன்மொழியப்பட்ட வேர்ச் சொற்களஞ்சியத்தில் காணப்பட்டால் பின்னொட்டு மற்றும் வேர்க் கோப்புகளில் இருந்து அவற்றின் தொடர்புடைய பொருள் தரப்படும். இது வேர் மற்றும் பின்னொட்டுக் கோப்புகளைச் சொல்லடைவு செய்வதன் மூலமும் செய்யப்படும். சொல்லடைவு செய்யப்படும் சமயத்தில் வேர் மற்றும் பின்னொட்டுச் சொற்களஞ்சியங்களுக்குச் சிறு அட்டவணைகள் (hash tables) உருவாக்கப்படும். வழிமுறை வரைவுகள் (algorithms) தொடர்புடைய சிறு அட்டவணைகளில் தொடர்புடைய சொல்லடைவுகளைப் பொருத்திப் பார்க்க வேண்டி பரிசோதிக்கும். முழு உள்ளீடும் தீர்க்கப்படுவது வரை இச்செயற்பாங்கு நிற்பதில்லை என்பதை முக்கியமாகக் கவனிக்க வேண்டும். இப்படி அது ஒரு பொருத்தத்தைக் கண்டுபிடித்து வெளியீட்டைத் தந்தாலும், சாத்தியம் என்றால் பகுப்பாய்வி கூடுதலான பகுப்புகளைக் கண்டு பிடிக்க இயலும் என்பதாகும். இது வேர் மற்றும் மாற்றுருபுகளுக்கு ஒரே ஒலியனியல் வடிவு இருந்தால் பகுப்பாய்வு இரண்டிற்கும் எல்லாச் சாத்தியமான பகுப்பாய்வையும் தரும். தற்போதைய நடைமுறைப்படுத்தலில் விளைவு (effective) வேறுபாட்டைக் காலப் பின்னொட்டில் பார்க்கப்படும் நிலையில் தமிழில் வினைகளுக்கு ஒரு தனி வேர் உள்ளீட்டைத் தருவது, ஜி.எஸ். மார்ஃபில் வேர் மற்றும் பின்னொட்டுக் கோப்புகள் அமைக்கப்படும் முறையின் காரணமாக இது சாத்தியமாகிறது. இது பகுப்பாய்வின் போது ஒரே நடப்பை மிக எளிதாகப் பொருண்மை மயக்கம் தீர்க்க உதவுகிறது.

ஜி.எஸ் மார்ஃப் பகுப்பாய்வி சொல்லின் வரிவடிவ உருபடுத்தத்தை உள்ளீடாக ஏற்கின்றது. சொற்கள் இடது எல்லையிலிருந்து பகுப்பாயப்படுகின்றன. முன்னர் விளக்கப் பட்டதுபோல் வேர் மற்றும் பின்னொட்டு சொற்களஞ்சியங்களின் பயன்பாட்டால் பகுப்பாய்வு வெற்றிகரமாக நடைபெறும். ஜி.எஸ். மார்ஃபின் உருவாக்கி உட்கூறு பகுப்பாய்வுக்கான அதே சொற்களஞ்சியங்களைப் பயன்படுத்துகின்றது. உள்ளீடு தரப்படும்போது உருவாக்கி அதை வேர்ச் சொற்களஞ்சியத்தில் பரிசோதித்து அதற்குத் தரப்பட்ட சொல்லடைவை எடுக்கும். அது பின்னர் அதே சொல்;லடைவைக் கொண்ட கோர்வைகளைப் பின்னொட்டு அகராதியில் தேடும். ஏதாவது கோர்வை அதே சொல்லடைவுடன் இருந்தால் வோ;மாற்றுரு அப்பின்னொட்டுக் கோர்வையுடன் சேரும். இது பின்னொட்டில் உள்ள உருபன்களின் அர்த்தங்களுடன் திருப்பி அனுப்பப்படும். பகுப்பாய்வுக்குப் பயன்படுத்தப்படும் அதே கோப்புகள் பயன்படுத்தப்படுவதால் உருவாக்குதலும் துல்லியமாக இருக்கின்றது (Winston Cruz, 2002:127)..

வின்ஸ்டன் குருஸ் (Winston Cruz, 2000) தமிழ் வினைச்சொற்களைப் பகுப்பாய்வு செய்வதற்கு ஜி.எஸ். மார்ஃப் முறையைக் கையாண்டார். ஜி.எஸ் மார்ஃப் என்பது உருபன்

அமைப்பொழுங்கு குறியீட்டால் ஆனது. வழிமுறை வரைவு (algorithm) இரண்டு கோப்புகளை பொருத்தமானதா இல்லையா என்று மட்டுமே ஆய்வு செய்கிறது. சொல்லாய்வி இரண்டு வடிவங்களை மட்டுமே பயன்படுத்தி இரண்டு கோப்புகள் படைக்கப்பட்டுள்ளது.

4.4.8. துரைபாண்டியின் உருபனியல் பகுப்பாய்வு

தற்கால தமிழில் வினைச்சொல் வடிவத்தினை உருவாக்கும் பகுப்பாய்வு இயந்திரத்தை முழுமையாக உருவாக்கியுள்ளார். இது இவர் உருவாக்கிய சொல்லாய்வியில்/சொல்லாளரில் பயன்படுத்தப் பட்டுள்ளது.

4.4.9. RCILTS-T-இன் தமிழ் உருபனியல் பகுப்பாய்வி

RCILTS-T (Resource Centre for Indian Language Technological Solutions-Tamil) என்ற அமைப்பின் தமிழ் உருபனியல் பகுப்பாய்வி அட்சரம் (atcharam) என அழைக்கப்படுகின்து. அட்சரம் வேர்ச் சொல்லிலிருந்து ஆக்கச் சொல்லைத் தனியாகப் பகுப்பாய்வு செய்து உருபனுடன் தொடர்புபடுத்துகிறது. இதனைப் பயன்படுத்தி 15 வகைகளை அடிப்படையாகக் கொண்டு இருபதாயிரம் வேர்ச்சொற்கள் தமிழ் அகராதியில் உருவாக்கப்பட்டுள்ளன. இது இரண்டு தொகுதிகளை உடையது. அதாவது பெயர் மற்றும் வினைச் சொல்லை அடிப்படையாகக் கொண்டது. இது ஊக விதிகளுடன் (heuristic rules) பொருண்மை மயக்கம் உடையதாகும். அதற்கு வினை மற்றும் பெயர் திரிபுகளைக் கையாண்டுள்ளனர்.

4.4.10. RCILTS-T-இன் உருபனியல் உருவாக்கி

இந்நிறுவனம் தமிழ்மொழியலும் உருபனியல் உருவாக்கியைத் தயார் செய்துள்ளது. அதற்கு அட்சயம் (Atchayam) எனப் பெயரிடப்பட்டுள்ளது. அட்சயம் என்பது பெயர் மற்றும் வினைச்சொற்களைக் கொண்டு இரண்டு முக்கிய உள்ளீடுகளைப் படைத்துள்ளது. இதில் பெயர்ச்சொல் பிரிவில் பின்னொட்டு இடம்பெறும். எடுத்துக்காட்டாக, பன்மை ஒட்டுக்களையும் திரிபு வடிவங்களையும் வேற்றுமை உருபுகளையும் பின்னுருபுகளையும் உருவாக்கியுள்ளனர். இதே போல் வினைப்பிரிவில் காலங்களையும் இடம்-எண்-பால் உருவாக்கிகளையும் பெயரச்சம், வினையெச்சம், பின்னொட்டு மற்றும் துணை வினைச்சொற்களையும் கொண்டுள்ளது. இது சந்தி விதிகள் மற்றும் 125 உருபனியல் விதிகளையும் பயன்படுத்தி உள்ளது. இது பெயரடை மற்றும் வினையடையையும் கையாண்டுள்ளது. இதில் சொல் மற்றும் வாக்கிய உருவாக்கியைக் கொண்டு ஆய்வு செய்யப்பட்டுள்ளது.

4.4.11. பொருண்மை மயக்கம் அல்லாத தமிழ் உருபனியல்

சொல் வடிவில் பொருண்மை மயக்கம் இருக்கும். இதற்கு மாறாக ஒப்புருச் சொல்லின் (homonymy) வகைப்பாடுகளில் எதிர் பாராத குறுக்கீடுகள் உருவாவதால் உருபனியல் பிறழ்ச்சி, ஒட்டுக்களின் பல்வகைச் செயல்பாடுகள் அல்லது நிச்சயமற்ற பின்னொட்டுச் சொற்கட்டுகளையும் (word bounding) உருவாக்குகிறது. இவ்விதம் வாக்கிய அமைப்புச் சூழல்களை இயல்பாகத் தீர்மானிக்கின்ற பகுப்பாய்வு ஆகும். இதனையே பொருண்மை மயக்கம் அல்லாத உருபனியல்

என்கிறோம். இரண்டு அடிப்படை அணுகுமுறைகள் சமவாய்ப்பு (probability) உடையது. விதியை அடிப்படையாகக் கொண்ட அணுகுமுறையில் சில பொருண்மை மயக்கம் தீர்வு காணப்பட முடியாமல் உள்ளது. ஆனால் மிகச் சில தவறுகளைக் கொண்டுள்ளது. பொதுவாகப் புள்ளியியல் அடையாளப்படுத்திகள் முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளைத் தருகின்றன. ஆனால் அவற்றறுள் நிறைய தவறுகள் காணப்படுகின்றன. சமவாய்ப்புள்ள முறையில் 1980-இருந்து உருபனியலில் பொருண்மை மயக்கமற்ற பகுப்பாய்வே அதிகம் இடம் பிடித்துள்ளது. நிலையான புள்ளியியல் முறைகளின் பயன்பாடு முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளாக உள்ளது. 'Influence of Morphology in Word Sense Disambiguation for Tamil' (தமிழ் உருபனியலில் பொருண்மை மயக்கம் அல்லாத சொல் செல்வாக்கு உடையது). பரிசோதனைகளை மேற்கொள்ளும்பொழுது மேற்பார்வை மற்றும் பகுதி மேற்பார்வை அணுகுமுறையை மேற்கொண்டு உருபனியல் திரிபுகளின் பாதிப்புமுறைகளை பற்றியும் திடமான தீர்வான உருபனியலில் பொருண்மை மயக்கம்மற்ற பகுப்பாய்வை தமிழ்மொழியிலும் பிற திரிபு மொழிகளிலும் பயன்படுத்தலாம் எனத் தீர்வு கண்டுள்ளனர்.

4.5. உருபனியலும் முற்றுநிலை மாற்றிகளும் (Morphology and Finite State Transducers)

ஒரு சொல்லின் கூறுகளை ஆயும் செயல்பாடுதான் உருபனியல் பகுத்துக்குறித்தல் (Morpholocal Parsing) எனப்படும். எடுத்துக்காட்டு மரங்கள் (மரம்) N + (கள்) PL மரங்கள் என்பது மரம் மற்றும் கள் என்ற இரு உருபன்களாகப் பிரியும் என்பதைத் தெரிந்து கொள்ளும் செயல்பாடுதான் உருபனியல் பகுத்துக் குறித்தல் எனப்படும். ஒன்றை உள்ளீடாக எடுத்துக் கொண்டு ஒரு வித அமைப்பை அதற்குத் தருவதைப் பொதுவாகப் பகுத்துக் குறித்தல் (parsing) எனலாம். ஆட்சி செய்தி (information retriveal) மீட்பு பரப்பில் (domain) மரங்கள் என்பதிலிருந்து மரம் என்பதைப் பொருத்திப் பார்க்கும் (mapping) சிக்கலை பகுதியாக்கம் (stemming) என்பார்கள். உருபனியல் பாகுப்பாய்வு செய்தலோ பகுதியாக்கம் செய்தாலோ பன்மை ஒட்டுகளுக்கு மட்டுமன்றி வேறு பல ஒட்டுகளுக்கும் பயன்படும். எடுத்துக்காட்டாக *வந்தான்* என்பதை வ + ந்த் + ஆன் என்று பகுக்கலாம் *வ* என்பது வினைப் பகுதி; *ந்த்* என்பது இறந்ததால ஒட்டு *ஆன்* என்பது ஆண்பால் ஒருமை வினைமுற்று விகுதி என்று அடையாளம் காணப்படும். நாம் தமிழ்ப் பெயர்களின் எல்லாம் திரிபு மற்றும் ஆக்க வடிவுகளையும் அகராதியல் பட்டியலிட்டுத் தருவதில்லை. அவ்வாறு செய்வது சிக்கனமான முறையல்ல. ஏனென்றால் *கள்* என்ற பன்மைப் பின் ஒட்டு விளைவாக்கமுள்ளது (productive). இது எல்லாப் பெயர்களுடனும் சேர்ந்து பன்மைப் பெயர் வடிவங்களை உருவாக்கும் தன்மை உடையது. எனவே வினைகள் மற்றும் பெயர்களின் எல்லாத் திரிபு வடிவுகளையும் பட்டியலிடுவது திறமையற்ற செயல் (அதாவது சிக்கனமானதல்ல); மேலும் விளைவாக்கம் உள்ள பின்னொட்டுகள் புதிய சொற்களுடனும் பயன்படுத்தப்படும். (எடுத்துக்காட்டாக கணினி – கணினிகள்) புதிய சொற்கள் நாள்தோறும் உருவாக்கப்படுகின்றன. இவற்றிற்கெல்லாம் பன்மை வடிவுகளைப் பட்டியலிட்டுக் கொண்டே செல்ல இயலாது. மேலும் துருக்கிய மொழி (Turkish language) போன்ற சிக்கலான உருபனியல் அமைப்புள்ள மொழிகளின்

சொற்களின் எல்லா வடிவுகளையும் அகராதியில் சேகரித்து வைக்க இயலாது. உருபனியல் குறைந்த பொருள்தருகின்ற அலகுகளான உருபன்களால் சொற்கள் எவ்வாறு உருவாக்கப்படுகிறது என்ற பாடம் உருபனியலாகும். உருபன் பொருண்மையைக் கொண்டிருக்கின்ற குறைந்த மொழி அலகாகும். எடுத்துக்காட்டாக *கால்* என்ற சொல்லில் ஒரு உருபன் உள்ளது. *கால்கள்* என்ற சொல்லில் *கால்*, *கள்* என்ற இரு உருபன்கள் உள்ளன. உருபன்களைப் பகுதிகள் (stems) ஒட்டுகள் (affxes) என்ற இரண்டு பெரிய உருபுகளாகப் பிரிக்கலாம். இவற்றிற்கு இடையே உள்ள வேறுபாடு மொழிக்கு மொழி வேறுபாடும் உள் உணர்வின் அடிப்படையில் பகுதி என்பது முக்கியப் பொருள் தரும் சொல்லின் முக்கியமான உருபன் என்றும் ஒட்டுகள் என்பது வேறுபட்ட வகையிலான கூடுதல் பொருள் ஊட்டும் உருபன்கள் என்றும் கூறலாம். ஒட்டுகள் முன்னொட்டுகள் (prefixes) பின்னொட்டுகள் (suffixes) இடையொட்டுகள் (infixes) மற்றும் ஓர ஒட்டுகள் (circumfixes) என்று பகுக்கலாம். முன்னொட்டுகள் பகுதிக்கு முன்வரும் பின்னொட்டுகள் பகுதிக்கு பின்னுமாக வரும். இடையொட்டுகள் பகுதிக்குள் சொருகப்படும். எடுத்துக்காட்டாக தமிழில் மான்கள் என்ற சொல்லில் வரும் பன்மை ஒட்டான (உருபன்) கள் பின்னொட்டாகும். அநீதி முன்னொட்டாகும். அநியாயம் என்ற சொற்களில் வரும் அ என்பது முன்னொட்டாகும். தமிழிலும் ஆங்கிலத்திலும் ஓர ஒட்டுகள் இல்லை. ஜெர்மன் மொழியில் ஓர ஒட்டுகள் உண்டு என்று பார்த்தோம்.

எடுத்துக்காட்டு:

film- en ‘ to film’ – ge-film-t ‘ filemed’

frag-en ‘to ask’ _ ge-frag-t ‘asked’

lob-en ‘to praise’ _ ge-lob-t ‘Praisd’

Zeig-en to ‘Show’ _ ge-zeig-en ‘Shown’

ge-யும் to-யும் இணைந்து Past Participle (இறந்தகால எச்சம்) வடிவை உருவாக்கும். பிலிப்பைன் மொழியான டாகலாக் (Tagalog) வியட்நாம் மொழியான சரவ் (Charau) போன்ற மொழிகளில் இடை ஒட்டுகள் பயன்படுத்தப்படுகின்றன.

சரவ் மொழியிலிருந்து எடுத்துக்காட்டு:

sulat ‘Write’ – s-um-ulat ‘wrote’

s- in-ulat ‘written’

டாகலாக் மொழியிலிருந்து எடுத்துக்காட்டு:

hingi ‘borrow’ – h-um-ingi

ஒரு சொல் பல உருபன்களின் பிணைக்கப்பட்டுக் உருவாக்கப்படுவதால் முன் ஒட்டுகளும், பின்னொட்டுகளும் பிணைப்பு உருபனியல் (concatenative morphology) எனப்படும். பல வழிகளில் உருபன்கள் மிகச் சிக்கலான வழியின் இணையும் பிணைப்பில்லாத உருபனியல் (non-

concutenative morphology) பரவலாக இருக்கும். மேற்சொன்ன டாகலாக் மொழி பிணைப்பு உருபனியலுக்கு எடுத்துக்காட்டாகும். வேறு ஒருவகையான பிணைப்பில்லாத உருபனியல் வார்ப்புச் சட்ட உருபனியல் (templatic morphology) அல்லது வேர் அமைப்பொழுங்கு உருபனியல் (root-pattern morphology) எனப்படும். அரேபிய மொழி (Arabic Language) ஹீப்ரூ மொழி பிற செமிட்டிக் மொழிகளில் (Semitic Language) சாதாரணமாக வார்ப்புச் சட்ட உருபனியல் காணப்படுகிறது. ஹீப்ரூ மொழியில் வினை இரண்டு உட்கூறுகளால் உருவாக்கப்படுகின்றது. ஒன்று அடிப்படைப் பொருளைக் கொண்டிருக்கும் மூன்று மெய்களால் (CCC) ஆன வேர் (வார்புச் சட்டம் இரண்டு பொதுவாக வினைப்பாட்டுப் (voice) பொருளை வெளிப்படுத்தும் உயிர்கள் எடுத்துக்காட்டாக மூன்று மெய்யால் ஆன 'learn' அல்லது 'study' என்று பொருள் தருகின்ற Lnad என்ற வேர் செய்வினையைக் (active voice) குறிக்கும் CaCaC என்ற வார்புச் சட்டத்துடன் சேர்ந்து *lamad* 'he studied' 'அவன் படித்தான்' என்ற சொல்லை உருவாக்கும் அல்லது வலியுறுத்தலைக் குறிக்கும். CiCeC என்ற வார்புச் சட்டத்துடன் சேர்ந்து *limed* 'he taught' 'அவன் கற்பித்தான்' என்ற சொல்லை உருவாக்கும் அல்லது வலியுறுத்தல் செயப்பாட்டு வினையைக் குறிக்கும். CuCaC என்ற வார்புச் சட்டத்துடன் சேர்ந்து *lumad* 'he was taught' அவன் கற்பிக்கப்பட்டான் என்ற சொல்லை உருவாக்கும். ஒரு சொல்லில் ஒன்றுக்கு மேற்பட்ட ஒட்டுக்கள் வரலாம். எடுத்துக்காட்டாக *பானைகளை* என்ற சொல்லில் *பானை* என்ற பகுதியும் *கள்* என்ற பன்மை விகுதியும் *ஐ* என்ற 2-ஆம் வேற்றுமை உருபும் உள்ளன. *rewrites* என்ற சொல்லல் *re* முன்னொட்டும் *write* என்ற பகுதியும் s என்ற பின்னொட்டும் உள்ளன. *unbelievably* என்ற ஆங்கிலச் சொல்லில் *un* என்ற முன்னொட்டும் *believe* என்ற பகுதியும் *able*, ly, என்ற பின்னொட்டுக்களும் உள்ளன. தமிழ் போன்ற மொழிகளில் மூன்றோ நான்கோ ஒட்டுக்களை கொண்டிருக்கும். ஆனால் துருக்கிய மொழி போன்ற மொழிகளில் சொற்கள் ஒன்பதோ பத்தோ ஒட்டுக்களைக் கொண்டிருக்கும். இம்மாதிரியான மொழிகள் ஒட்டுக்கள் நிறைந்த மொழிகள் (agglutinative language) என்று அழைக்கப்படும். உருபன்களிலிருந்து சொற்கள் உருவாக்கப்படுவதை இரண்டு பெரும் வகுப்புகளாகப் பகுக்கலாம். அவைகள் திரிபுகள் (inflection) ஆக்கங்கள் (derivation) என்பன. ஒரு சொல் பகுதி ஒரு இலக்கண உருபனுடன் இணைந்து அதே சொல்வகுப்பைச் சேர்ந்த சொல்வடிவை விளைவிப்பது திரிபு எனப்படும். எடுத்துக்காட்டாக பன்மை ஒட்டான *கள்* ஒருமைப் பெயர்ப்பகுதியுடன் இணைந்து பன்மை இலக்கண உருபனுடன் இணைந்து பொதுவாக வேறு ஒரு வகுப்பைச் சார்ந்த பெரும்பாலும் பொருளைச் சிக்கலான சொல்வடிவை உருவாக்குவது ஆக்கம் (derivation) எனப்படும்.

4.5.1 முற்றுநிலை உருபனியல் பகுத்துக்குறித்தல் (Finite State Morphological Parsing)

முற்றுநிலைத் தானியங்யின் உருபனியல் பகுத்திக்குறித்தலைப் பின்வரும் எடுத்துக்காட்டு மூலம் விளக்கலாம். நமது நோக்கம் முதல் செங்குத்துப் பத்தியில் கொடுத்துள்ள உள்ளீட்டுச் சொற்களுக்கு இரண்டாவது பத்தியில் தரப்பட்டுள்ள வெளியீட்டைப் பெறுவதாகும்.

உள்ளீடு	**உருபனியல் பகுத்துக்குறித்தல் வெளியீடு**
கால்கள்	கால் + பெயர் + பன்மை
கால்	கால் + பெயர் + ஒருமை
கடாக்கள்	கடா + பெயர் + பன்மை
கடா	கடா +பெயர் + ஒருமை
மரங்கள்	மரம் + பெயர் + பன்மை
மரம்	மரம்+பெயர் + ஒருமை
கற்கள்	கல் + பெயர் + பன்மை
கல்	கல்+ பெயர் + ஒருமை
முட்கள்	முள் + பெயர் + பன்மை
முள்	முள்+ பெயர் + ஒருமை

முதல் குத்துநிலை வரிசையில் தரப்பட்ட வடிவங்களுக்கு இரண்டாவது குத்துநிலை வரிசையில் தரப்பட்டவாறு வெளியீட்டைப் பெறுதல் நமது நோக்கமாகும். இரண்டாவது வரிசையில் ஒவ்வொரு சொல்லின் பகுதியும் உருபனியல் பண்பு கூறுகளாகும். எடுத்துக்காட்டாக +பெயர் என்பது அந்த சொல் ஒரு பெயர் என்பதையும் +ஒருமை என்பது ஒருமை எண் என்பதையும் +பன்மை என்பது பன்மை எண் என்பதையும் குறிப்பிடும்.

உருபனியல் பகுப்பான் (Morphological Parser) உருவாக்குவதற்கு பின்வருவன தேவை.

1. அகராதி (Lexicon): பகுதிகள் மற்றும் ஒட்டுகள் இவற்றின் அடிப்படைச் செய்திகளுடன் கூடிய (அதாவது பத்தி பெயரா வினையா போன்ற செய்திகள்) அடங்கிய பட்டியல்

2. உருபன் அமைப்பொழுங்கு: ஒரு சொல்லுக்குள்ள எந்த வகையைச் சார்ந்த உருபன்கள் பிற வகுப்புகளைத் தொடர்;ந்து வரும் என்பதை விளக்குகின்றது. உருபன் வருகை முறையின் மாதிரி எடுத்துக்காட்டாக தமிழில் பன்மை உருபு பகுதியைத் தொடர்ந்து வரும். பகுதிக்கு முன்னர் வராது போன்ற செய்திகளைக் கூறலாம்.

மரம் + கள் +ஐ

- கள் - மரம் - ஐ
- ஐ – கள் - மரம்
- கள் - ஐ – மரம்
- மரம் - ஐ – கள்

3. சந்திவிதிகள்/உருபஒலியனியல் விதிகள் அல்லது எழுத்தலகு விதிகள் (Morphophonemic Rules அல்லது Orthographic Rules) தமிழில் சந்தி விதிகள் சாதாரணமாய் இரு உருபன்கள்

இணைகையில் ஏற்படும் மாற்றங்களை மாதிரிப்படுத்ததப் பயன்படுகின்றன. ஆங்கிலம் போன்ற மொழிகளில் எழுத்துக்கூட்டல் (Spellers) விதிகள் சாதாரணமாய் இரு உருபன்கள் இணைகையில் ஏற்படும் மாற்றங்களை மாதிரிப்படுத்தும்.

மரம் + பன்மை > மரங்கள்

மரம் + ஐ > மரத்தை

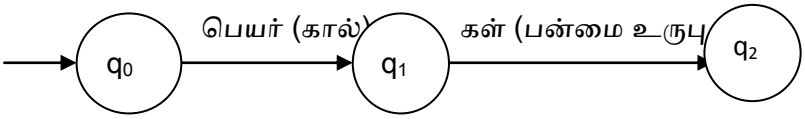
city + s > cities

box + s > boxes

உருபன் அமைப்பொழுங்கு பற்றிய அறிவை மாதிரிப்படுத்த எவ்வாறு முற்றுநிலைத் தானியங்கியைப் பயன்படுத்துவது உருபன்களைத் தெரிந்து கொள்ளலின் துணைச் சிக்கல்களை அகராதியில் எவ்வாறு எளிமையாக உருப்படுத்தம் செய்வது என்பது அடுத்து எதிர் கொள்ள வேண்டிய பிரச்சனைகள் ஆகும். இதற்கு நாம் முற்றுநிலை மாற்றிகள் (Finite State Transducers) அகராதியில் உருபனியல் பண்புக் கூறுகளை மாதிரிப்படுத்த அறிமுகப்படுத்த வேண்டி வரும். இறுதியில் எவ்வாறு எழுத்தலகு விதிகளை அல்லது உருபொலியனியல் விதிகளை மாதிரிப்படுத்த முற்றுநிலை மாறிகளை பயன்படுத்தலாம் என்பதையும் தர வேண்டும்.

4. 5.2 அகராதியும் உருபன் அமைப்பொழுபங்கும் (Lexicon and Morphotactics)

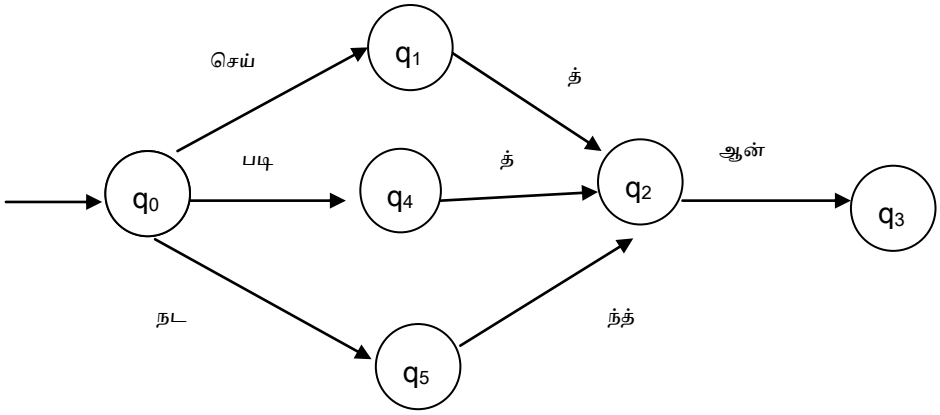
அகராதி சொற்களை வைப்பதற்கான ஒரு களஞ்சியம் எளிய அகராதி மொழியின் எல்லா சொற்களையும் இயற்பெயர்களையும் கொண்டிருக்கும். ஒரு மொழியில் உள்ள எல்லாச் சொற்களையும் பட்டியலிடுவது வசதியானதாகவோ இயலாமலோ இருப்பதால் கணினி அகராதி சாதாரணமாக மொழியின் ஒவ்வொரு பகுதியையும் ஒட்டுகளையும் அவை எவ்வாறு இணையும் என்ற செய்தியைத் தருகின்ற உருபன் அமைப்பொழுங்கின் உருப்படுத்தத்துடன் பட்டியலிட்டு கட்டமைப்பு செய்து தரும். உருபன் அமைப்பொழுங்கை பல வகைகளில் மாதிரிப்படுத்தலாம். மிகப் பொதுவான ஒரு மாதிரி முற்றுநிலைத் தானியங்கியாகும். பின் வருவது தமிழ்ப் பன்மை திரிபின் முற்றுநிலை மாதிரியாகும்.



பின்வருவது தமிழ் வினைத்திரிபின் முற்று நிலை மாதிரியாகும்

செய்-வகுப்பு வினைகள்	படி-வகுப்பு வினைகள்	நட-வகுப்பு வினைகள்	செய்-வகுப்பு வினைகளின் இறந்தகால ஒட்டு	படி-வகுப்பு வினைகளின் இறந்தகால ஒட்டு	நட-வகுப்பு வினைகளின் இறந்தகால ஒட்டு	படர்க்கை ஆண்பால் ஒட்டு
செய்	படி	மற	த்	த்த்	ந்த்	ஆன்

பெய்	பிடி குடி	நட அள				



முற்றுநிலைத் தானியங்கியை பயன்படுத்தி உருபனியல் தெரிந்துகொள்வதின் சிக்கலை தீர்க்கலாம். அதாவது உள்ளீடு செய்கின்ற எழுத்து சரியான தமிழ்ச் சொல்லா இல்லையா என்று உறுதிசெய்து கொள்ளலாம்.

4.5.3 முற்றுநிலை மாறிகளின் உதவியால் உருபனியல் பகுப்பாய்வு

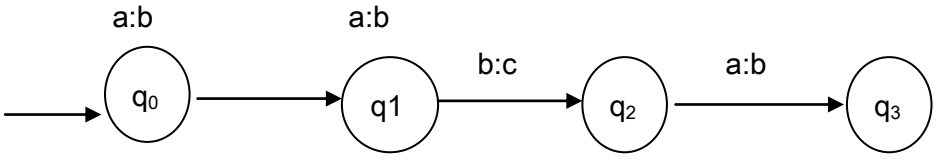
கொஸ்கெனிமி (Koskenniemi) எடுத்துச் சொன்ன இருநிலை உருபனியல் (Two Level Morphological) உருபனியல் பகுப்பாய்வைச் செய்யலாம். இருநிலை உருபனியல் ஒரு சொல்லை உருவாக்குகின்ற உருபன்களின் எளிய இணைப்பை உருப்படுத்தம் செய்யும் சொல்சார் நிலைக்கும் (Lexical Level) (அக அமைப்புக்கும்) இறுதிச் சொல் வடிவின் உண்மையான எழுத்துக் கூட்டலை உருப்படுத்தம் செய்யும் புற அமைப்புக்கும் இடையே உள்ள பொருத்தத்தை உருப்படுத்தம் செய்கின்றது. எழுத்துக் கோர்வைகளால் ஆன மரங்கள் போன்ற புற அமைப்பு நிலையை மரம் + பெயர் + பன்மை என்ற உருபன் பண்புக்கூறுகளால் ஆன தொடர்ச்சியாக சொல்சார் நிலையை (அக அமைப்பை) அறிவது உருபனியல் பகுப்பாய்வாகும்.

சொல்நிலை		வ	ண்	டு	N	PL	

புறநிலை		வ	ண்	டு	க	ள்	

சொன்மை மற்றும் புற நாடாக்களின் எடுத்துக்காட்டு இவ்வாறு இருநிலைக்களுக்கிடையே

உருபடுத்தத்தைச் செயல்படுத்த நாம் பயன்படுத்தும் தானியங்கி முற்றுநிலை மாற்றியாகும் (FST) மாற்றிகள் ஒரு குமுமக்குறியீடுகளுக்கும் மற்றொரு குழுமக்குறியீடுகளுக்கும் இடையே உருப்படுத்தம் செய்கின்றது. முற்றுநிலை மாற்றி இதை முற்றுநிலைத் தானியங்கி வழியாகச் செய்கின்றது. இவ்வாறு நாம் ஒரு முற்றுநிலை மாற்றியைக் கோர்வைகளின் இணைகளை தெரிந்துகொள்கின்ற அல்லது உருவாக்குகின்ற இரு நாடா முற்றுநிலை தானியங்கியைக் காணலாம். இவ்வாறு முற்றுநிலை மாற்றிக்கு முற்றுநிலை தானியங்கியை விட கூடுதல் பொதுமைச் செயால்பாடு இருக்கின்றது. முற்றுநிலை தானியங்கி ஒரு முறையான மொழியைக் கோர்வைகளின் ஒரு குழுமமாக விவரணை செய்கின்றது. ஆனால் முற்றுநிலை மாற்றி ஒரு முறையான மொழியைக் கோர்வைகளின் குழுமங்களுக்கு இடையே உள்ள உறவாக விவரணை செய்கின்றது. இது முற்றுநிலை மாறியை ஒரு கோர்வையை உருவாக்கும் ஒரு பொறியாகக் காட்டுகின்றது. ஒரு உள்ளீட்டாலும் வெளிட்டாலும் ஒவ்வொரு வில்லும் அடையாளப்படுத்தப்பட்ட முற்றுநிலைப்பொறி முற்றுநிலை மாற்றியாகும். ஒரு தொடர்பை முற்றுநிலைமாற்றியாக உருப்படுத்தம் செய்யலாம்.



இணைகளின் ஒரு முற்றுநிலை அல்லது எச்சநிலைக் குழுமம் ஒரு உறவாகும்.

R = (1,1), (2,4), (3,9), (4,16), (5,25)

S = (a, A), (b, B), (c, C), (d, D), (e, E)….

U = (கேள்வி i, பதில் i)

V = (ராமன் வீட்டுக்குப் போய்க்கொண்டிருக்கிறான் Raman is going home)

W = (a,b), (aaa,bbb), (aaaa,bbbb).....

ஒரு உறவை பார்வை அட்டவணையாக (look up table) (உள்ளீடு, வெளியீடு) உருப்படுத்தம் செய்யலாம். அந்த அட்டவணை எல்லையற்ற அளவு பெரிதாய் இருக்கலாம்; பின் தரப்பட்டுள்ளது மாற்றிகளின் நான்கு மடங்கு வழிகளின் சுருக்கமாகும்.

1. அறிவானாக முற்றுநிலைமாற்றி (FST as recognizers)

கோர்வைகளின் ஜோடியை/இணையை உள்ளீடாகவும் வெளியீடாகவும் எடுத்துக்கொண்டு கோர்வை இணைமொழியின் கோர்வை இணையாக இருந்தால் ஏற்றுக்கொள்ளவும் இல்லாவிட்டால் விட்டுவிடவும் செய்யும் மாற்றி முற்றுநிலைமாற்றி அறிவானாகும்.

2. உருவாக்கியாக முற்றுநிலைமாற்றி (FST as generator)

மொழியின் கோர்வைகளின் இணைகளை வெளியீடு செய்கிற பொறி முற்றுநிலை மாற்றி உருவாக்கியாகும் இவ்வாறு வெளியீடு "ஆம்" அல்லது இல்லை என்பாதம் மற்றும் வெளியீட்டுக் கோர்வைகளின் இணையாகும்.

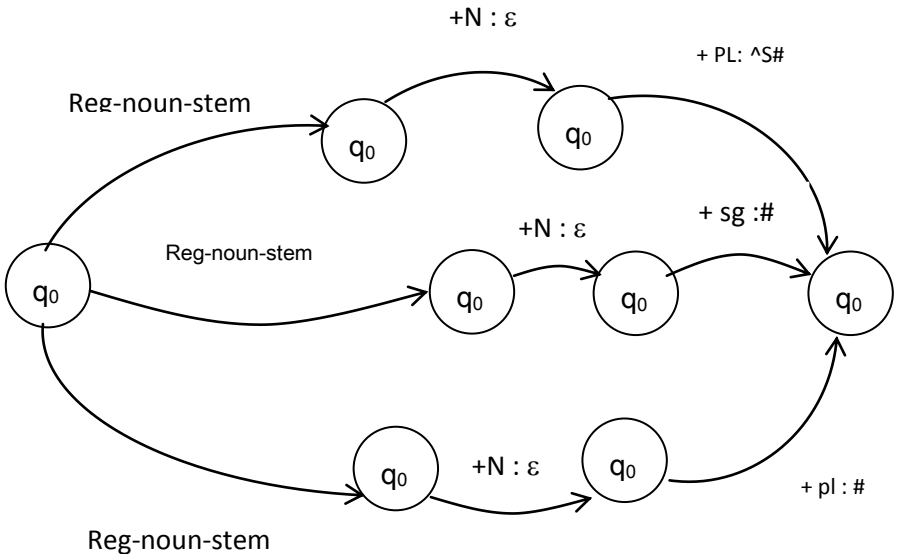
3. மொழிபெயர்ப்பியாக முற்றுநிலை மாற்றி (FST as translator)

ஒரு கோர்வையைப் படித்து மற்றொரு கோர்வையை வெளியீடு செய்கிற பொறி முற்றுநிலைமாற்றி மொழிபெயர்ப்பியாகும்.

4. உறவுபடுத்தியாக முற்றுநிலைமாற்றி (FST as realtor)

குழுமங்களுக்கு இடையே உறவுகளைக் கணிக்கிற பொறி முற்றுநிலைமாற்றி உறவுப்படுத்தியாகும். எங்கு முற்றுநிலைத்தானியங்கிகள் முறையான மொழிகளுடன் சம உருவம் கொண்டிருக்கிறதோ அங்கு முற்றுநிலைமாற்றிகள் முறையான உறவுகளுடன் சமஉருவம் கொண்டிருக்கும். கோர்வைகளின் குழுமங்களாக இருக்கிற முறையான மொழிகளின் இயற்கையின் நீ;ட்சியான கோர்வைகளின் இணைகளின் குழுமங்கள் முறையான உறவுகளாகும். எ.கா ஆங்கில பன்மைத் திரிபு அகராதியில் கலவை நிலையில் அமையும்.

reg – noun	irreg – pl – noun	irreg –sg -noun
fox	g o:e o:e s e	goose
cat	sheep	sheep
dog	m o:i u:ɛ s:c e	mouse



ஆங்கில பெயர்திரிபின் மாற்றி q_1, q_2 என்பன ஏற்றுக்கொள்ளும் நிலைகள் ஆகையால், சீரான/முறையான பெயர்கள் பன்மைப் பின்னொட்டு பெறவோ பெறாமலோ இருக்கலாம். ^ என்பது உருபன் - எல்லைக் குறியீடு: # என்பது சொல் எல்லைக் குறியீடு

சொல்நிலை	F	O	X	+N	+PL

இடைப்பட்டநிலை	f	o	x	^	s	#

4.5.4 எழுத்தலகு விதிகளும் முற்று நிலை மாற்றிகளும்

ஆங்கிலத்திற்கு உருபன் எல்லையில் எழுத்துக் கூட்டல் மாற்றங்கள் தேவை. அதற்காக நாம் எழுத்துக் கூட்டல் விதிகள் அல்லது எழுத்தலகு விதிகளை அறிமுகம் செய்ய வேண்டிவரும். நாம் எழுத்துக் கூட்டல் மாற்றங்களை எளிய உருபன்களின் இணைப்பின் உள்ளீட்டாகவும் கொஞ்சம் மாறுபட்ட உருபன்களின் இணைப்பின் வெளியீடாகவும் எடுத்துக்கொள்ளலாம். பின்வருவது மூன்று நிலைகளைக் காட்டுகின்றது: சொல் நிலை, இடைநிலை, புறநிலை.

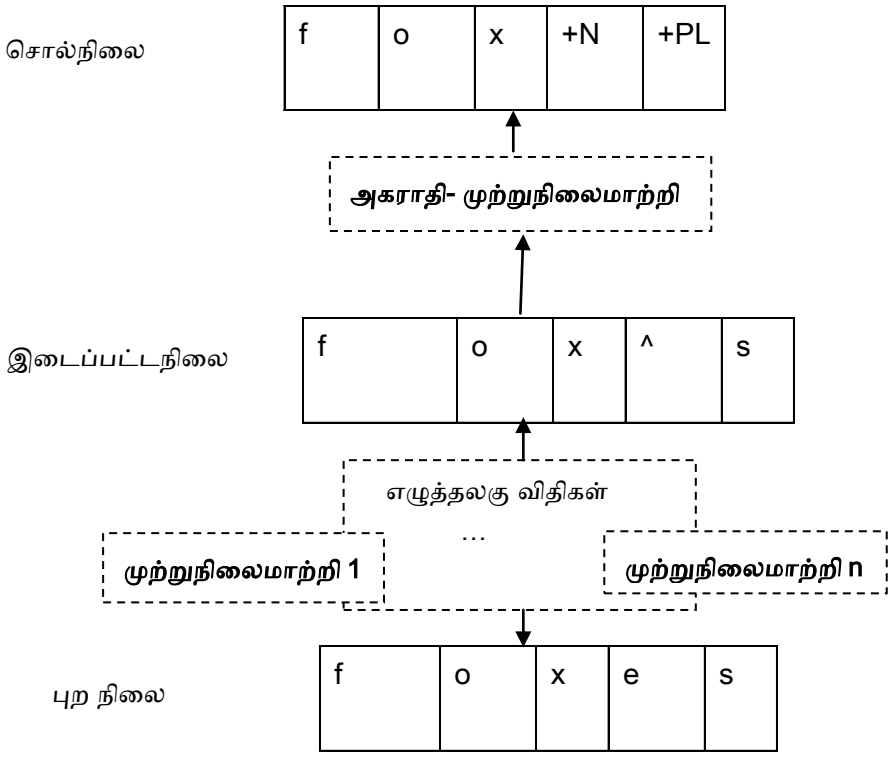
சொல்நிலை	f	o	x	+N	+pl

இடைப்பட்டநிலை	f	o	x	^	s	#

புற நிலை	f	o	x	e	s	#

4.5.5 .முற்றுநிலை மாற்றியின் அகராதியையும் விதிகளையும் சேர்த்தல் (Combining FST Lexicon and rules)

நாம் பகுத்துக்குறிக்கவும் உருவாக்கவும் அகராதியையும் விதி மாற்றிகளையும் இணைக்க தயாராகிவிட்டோம். பின்வரும் படம் பகுத்துக் குறிக்கவோ உருவாக்கவோ பயன்படும் இருநிலை உருபன் ஒழுங்குமுறையின் அமைவைக் காட்டுகின்றது.



அகராதி மாற்றி அதன் பகுதிகள் மற்றும் உருபன் பண்புக்கூறுகள் இவற்றைக் கொண்டு சொல் நிலைக்கும் உருபன்களை இணைவதை உருப்படுத்தம் செய்யும் இடைநிலைக்கும் இடையே பொருத்தம் செய்கின்றது. ஒவ்வொரு தனி எழுத்துக் கூட்டல் விதிக்கட்டுப்பாட்டையும் உருப்படுத்தம் செய்கின்ற ஒரு கூட்டம் மாற்றிகள் இடைநிலைக்கும் புறநிலைக்கும் இடையே பொருத்தம் செய்கின்றது.

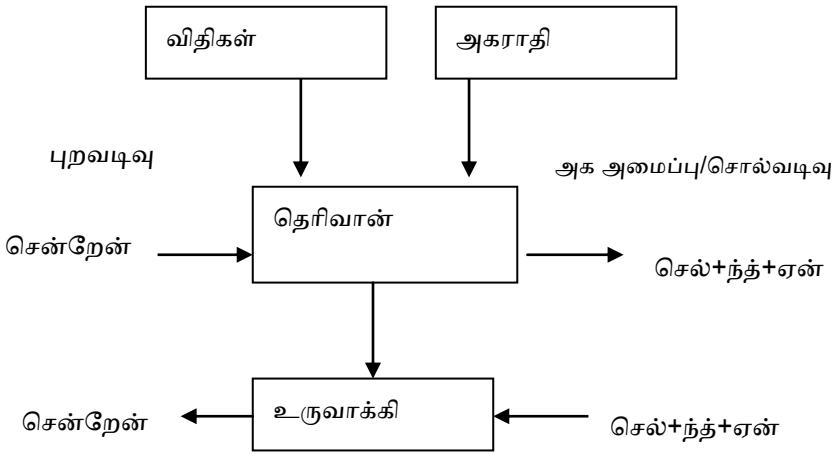
4.5.6. பிசிகிமோ (Pckimmo)

கொசொகென்னிமியின் (Kosekenniemi) இருநிலை உருபனியல் பரவலாகப்; பயன்படுத்தப்படுகின்ற குறிப்பாக ஐரோப்பிய மொழிகளில் பயன்படுத்தப்படுகின்ற ஒன்று. பிசிகிமோ உருபனியல் ஆய்வியாக இரண்டு செயல்களைச் செய்ய இயலும்.

1. தெரிந்துக்கொள்ளுதல் (recognition)
2. உருவாக்குதல் (generation)

பிசிகிமோவின் உருவாக்கும் பகுதி வேரின் சொல்வடிவை உள்ளீடாக ஏற்கும். தெரிவான்/உணரி புற அமைப்பை உள்ளீடாக ஏற்கும் விதிகளை பயன்படுத்தும்: அகரதியைப் பார்க்கும் தொடர்புள்ள சொல்வடிவையும் பொருள் கோர்வையும் தரும். இதைப் பயன்படுத்துவதில் உள்ள

எளிமையும் இம்மாதிரியின் சிக்கனம் அடிப்படையிலான முக்கியத்துவமும் என்னவென்றால் ஒரே குழும விதிகளை தெரிந்துகொள்ளும் செயல்பாட்டிற்கும் உருவாக்கும் செயல்பாட்டிற்கும் பயன்படுத்தலாம். விதிகளை இருதிசைகளிலும் பயன்படுத்தலாம். சில ஒலியன் விதிகளோ உருபன் விதிகளோ எழுத்தலகு விதிகளோ புற அமைப்பிலிருந்து, அக அமைப்பைத் தரும் மாற்றமாகக் கொள்ளப்பட்டால் அவற்றைத் திருப்பிச் செயல்பாட்டிற்காக (சுநஎநசளந pசடிஉநளள) மீண்டும் எழுதத் தேவையில்லை. திரும்புகையில் அது புற அமைப்பிலிருந்து அகஅமைப்பை ஆய்கிறது. இதைப் பின்வரும் படத்தில் காட்டலாம்.



இம்மாதிரியில் பயன்படுத்தப்படும் இருநிலைகள் எல்லாம் மொழியியல் விவரணைகளாகும். இந்த விவரணைகள் மாற்றொலியன் பிறழ்ச்சிகள் (allophonic alternations) உருப ஒலியன் பிறழ்ச்சிகள் (morphophonemic alternations)அல்லது எழுத்தலகு வேறுபாடுகள் (orthographic variations) இவற்றை கையாள இயலும். பிசிகிமோ இருநிலை மொழியியல் விவரணைகளை முற்றுநிலை அட்டவணைகளாக மாற்றும். விரிவாகக்கூறப் புகுந்தால் விதிக்கோப்பில் விதிகள் முற்றுநிலை மாற்றிகளாக (FST) உருபடுத்தம் செய்யப்படும் உருபன் அமைப்பொழுங்குக் கட்டுப்பாடு (Morphotatic Constraints) அகராதிக் கோப்பில் முற்றநிலைத் தானியங்கியாகக் (FST)கூறப்படும்.

அகராதிப் பகுதி (Lexical Component)

பிசிகிமோவின் அகராதிப் பகுதி தெரிந்துகொள்ளும் செயல்பாட்டின் போது மட்டும்தான் நோக்கப்படும்; உருவாக்கும் செயல்பாட்டில் நோக்கப்படமாட்டாது. அகராதி ஒரு சொல்லின் உருப்படுத்தத்தைக் காட்டும் அதன் உருபன் அமைப்பொழுங்குக் கட்டுப்பாட்டைக் குறிப்பிடும். ஒவ்வொரு சொல்லிற்குப் பொருள் தரும் பிசிகிமோவின் ஒவ்வொரு அகராதியும் துணை அகராதிகளாகப் பகுக்கப்பட்டிருக்கும். ஒவ்வொரு துணை அகராதியும் உருபன் அமைப்பொழுங்கில் ஒரே மாதிரி செயலாற்றுகிற ஒரு வகுப்பைச் சார்;ந்த சொற்களை உருப்படுத்தம் செய்யும். ஒவ்வொரு சொற்பதிவும் (lexical entry) ஒரு சொல்லையும் அதன்

தொடர்ச்சி வகுப்பையும் அதன் பொருளையும் கொண்டிருக்கும். தொடர்ச்சி வகுப்பு அதைத் தொடர்ந்து வரும் வேறுபட்ட உருபன்களையும் இலக்கணத் தன்மையின் உருபன் அமைப்பொழுங்கு விதிகள் எங்கு கூறப்பட வேண்டும் என்பதைக் காட்டும். ஒரு சொல்லைத் தொடர்ந்து வரும் துணை அகராதிக் குழுவை விவரிக்கும் பிறழ்ச்சியின் பட்டியல் இருக்கும். எடுத்துக்காட்டாக, *கால்* என்ற சொல்லை எடுத்துக் கொள்வோம் அது *ஐ* என்ற இரண்டாம் வேற்றுமை உருபையும் *உக்கு* என்ற நான்காம் வேற்றுமை உருபையும் *இல்* என்ற இட வேற்றுமை உருருபையும் இலிருந்து என்ற நீக்கல் வேற்றுமை உருபையும் கள் என்ற பன்மை உருபையும் ஆ என்ற வினா பின்னொட்டையும் ஏற்கும். சேய்மை கட்டு அடையில் *அ* என்பது அந்த என்பதன் மாற்று உருபாகும்.

4.6. சுருக்கம்

இவ்வியல் கணினி உருபனியல், தமிழ் உருபனியல், உருபனியலும் முற்றுநிலை மாறிகளும் என நன்கு தலைப்புகளில் அமைகின்றது. முதலில் கணினி உருபனியல் அறிமுகப்படுத்தப்பட்டு உருபனியலை மாதிரிப்படுத்தலில் சில உருபனியல் மாதிரிகள் விளக்கப்பட்டுள்ளது. இரண்டாவதாக தமிழ் உருபனியல் சுருக்கமாக விளக்கப்பட்டுள்ளது. மூன்றாவது தமிழுக்காக உருவாக்கப்பட்ட சில முக்கியமான உருபனியல் பகுப்பாய்விகள் பட்டியலிடிப்பட்டு அவற்றைப் பற்றி சுருக்கமான தகவல்கள் தரப்பட்டுள்ளன. நான்காவதாக முற்றுநிலை மாற்றிகள் அறிமுகப்படுத்தப்பட்டு அவற்றைப் பயன்படுதி உருபனியல் பகுப்பாய்வும் உருவாக்கமும் செய்யும் செயலாக்கம் விளக்கப்பட்டுள்ளது. இறுதியில் உருபனியல் ஆய்வு, உருவாக்கம் என்ற இரண்டையும் நடைமுறைபடுத்தும் பிசிகிமொவின் அமைப்பும் செயல்பாடும் விளக்கப்பட்டுள்ளது.

இயல் 5

கணினித் தொடரியல்

5.1. அறிமுகம்

கணினி தொடரியல் (computational syntax) ஆய்வு வாக்கியங்களைத் தொடர்களாகவும் சொற்களாகவும் பகுத்து அவற்றின் தொடரியல் மற்றும் சொல்வகைப்பாட்டுப் பண்புக்கூறுகளை அடையாளப்படுத்தித் தரும் செயல்பாடாகும். வாக்கியங்கள் தரப்படுகையில் அவற்றை தொடரியல் கூறுகளாகப் பகுப்பது மற்றும் தொடரியல் கூறுகள் தரப்படும் போது வாக்கியங்களை உருவாக்குவது என்ற இரு செயல்பாடுகள் இதில் அடங்கும்.

5.2. கணிப்பொறி வழி மொழி ஆய்வுக்கான இலக்கண வடிவமைப்புகள்

ஒரு கூற்றின் பொருத்தமான பொருளைத் தருமாறு சொற்களுக்கு இடையே உள்ள உறவுகளைப் பிரித்தெடுக்கும் நுட்பத்தைத் தருவது இந்த இலக்கண வடிவமைப்பின் முக்கியமான செயல்பாடாகும். வாக்கியங்களை பகுப்பாய்வு செய்வதற்குப் பயன்படுத்துகின்ற செயல்முறையிலிருந்து தனியாக ஒரு குழும விதிகளால் நாம் மொழியை பண்பாக்கம் செய்ய விரும்புகிறோம். அம்மாதிரியான குழும விதிகள் முறையான இலக்கணம் எனப்படும். முறையான இலக்கணம் ஒரு குழும இலக்கண வாக்கியங்களை வரையறுக்கிறது. அவ்வாக்கியங்களுக்கு அமைப்பைத் தருகின்றது. மொழியைப் பேசுகின்றவர்களின் இலக்கணம் குறித்த உள்ளுணர்வுடன் பொருத்துகின்ற வாக்கியத்தின் பொருளை வரையறுக்கின்ற முறையான இலக்கணத்தை உருவாக்குகிறதுதான் நமது செயல். தொடக்ககால கட்டத்தில் சூழல்வரையறை இல்லாத இலக்கண வடிவவாதம் பயன்படுத்தப்பட்டது. தற்போது கட்டுப்பாடு அடிப்படையிலான இலக்கண முன்மாதிரிகள் பயன்படுத்தப்படுகின்றன. அவை கீழே பட்டியலிடப்பட்டுள்ளன:

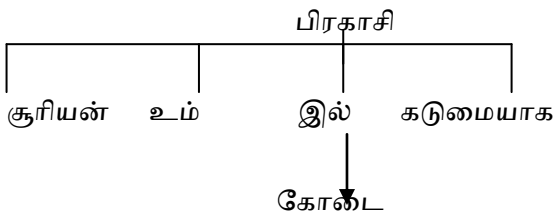
- செயல்பாட்டு ஒருங்கிணைக்கப்பட்ட இலக்கணம் (Functional Unification Grammar)
- தலை இயக்கத் தொடரமைப்பு இலக்கணம் (Head-driven phrase Structure Grmmar)
- செயல்பாட்டு இலக்கணம் (Fuctional Grammar)
- சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar)
- வகைப்பாட்டு ஒன்றிணைக்கப்பட்ட இலக்கணம் (Categorical Unification Grammar)
- கிளை சேர்க்கப்படும் இலக்கணம் (Tree adjoining Grammar)
- சார்பு இலக்கணம் (Dependency grammar)

முறையான இலக்கணம் (Formal Grammar)

வாக்கியங்களைப் பகுப்பாய்வு செய்வதற்குப் பயன்படுத்துகின்ற செயல்முறையிலிருந்து தனியாக ஒரு குழும விதிகளால் நாம் மொழியை பண்பாக்கம் செய்ய விரும்புகிறோம். அம்மாதிரியான குழும விதிகள் முறையான இலக்கணம் ஒரு குழும இலக்கண வாக்கியங்களை வரையறுக்கிறது; அந்த வாக்கியங்களுக்கு அமைப்பைத் தருகின்றது. மொழியைப் பேசுகின்றவர்களின் இலக்கணம் குறித்த உள்ளுணர்வுடன் பொருத்துகின்ற வாக்கியத்தின் பொருளை வரையறுக்கின்ற முறையான இலக்கணத்தை உருவாக்குகிறதுதான் நமது செயல்.

5.2.1 சூழல் வரையறை இல்லாத இலக்கணம் (Context Free Grammar)

சூழல் வரையறை இல்லாத இலக்கணம் மரபு மொழியியலரால் அண்மை உறுப்பு இலக்கணங்கள் என்றும் வழியமைப்பு மொழித் திட்டமிடுபவர்களால் பேக்ஸ் இயல்பான வடிவு (Bgackes Normal Form) என்றும் சில கணினி பயன்பாட்டில் மறுதரவு அமைப்பொழுங்கு (Recursive Pattern) என்றும் அழைக்கப்படுகின்றது. இவைகளைப் பொதுவாகத் தொடரமைப்பு இலக்கணம் (Pharse Structare Grammar) என்பதன் குறிப்பிட்ட வகை எனலாம்; இவைதான் ஆக்கமுறை இலக்கணத்திற்கு (Generative Grammar) அடிப்படையாக அமைந்தது. அமைப்புக்கான மூன்று அணுகுமுறைகள்: புடைப்பெயர்வு பின்னல் (Transition Network), தட்டையான அமைப்பு (flat structure) உடையவை, கிளையமைப்பு (hierarchical structure) உடையவை. நம்முடைய உள்ளுணர்வு உறுப்பமைவை எதிர்பார்க்கின்றது. இதை வாக்கியம் என்பது பின்னலால் விற்களின் தொடர்ச்சியாகப் பொருந்துவதாக கொள்ள இயலாது. வாக்கியம் துண்டுகளால் அல்லது தொடர்களால் ஆனது. தொடர்களின் அமைப்பு, பொருள் எவ்வாறு பெறப்படுகின்றது என்று கூறுவதால் இவ்வகையிலான அமைப்பாக்கம் முக்கியமானதாகும். மொழி அமைப்பை விளக்க பல வகைப்பட்ட விதிகளின் வடிவமைப்புகள் (formulations) இருக்கின்றன. எவ்வாறு மொழி அமைப்பை நன்றாக விளக்கலாம் என்ற அணுகுமுறையின் அடிப்படையில் அவற்றின் விளக்கங்கள் வேறுபடுகின்றன. தலை மற்றும் அடை அணுகுமுறை (head and modifier approach), இலக்கணத்தின் மரபுத்தன்மையிலான பல அறிமுகங்கங்கள் அமைப்புகளை அடிப்படை அமைப்பொழுங்கிலிருந்து நீட்சி செய்யப்பட்ட அடை செய்யப்பட்ட அமைப்பாக விவரணை செய்கின்றன. எடுத்துக்காட்டாக, வாக்கியத்தின் அக அமைப்புப் பெயர்த் தொடரைத் தொடர்ந்து வரும் வினை என்று கொள்ளலாம். பெயர் அடைகள், சுட்டடைகள் (அடைகொளி அடைகள்) என்பன பெயரை அடை செய்வதாகவும் வினையடை போன்றவை வினையை அடை செய்வதாகவும் கருதலாம்; அடைகளையும் அடைகளால் விவரிக்கலாம். ஒவ்வொறு நிலையிலும் தனிச்சொல்லாக ஒரு தலை இருக்கும். அதற்கு ஒன்றோ அதற்கு மேலோ அடைகள் வரலாம். எடுத்துக்காட்டாக, *சூரியன் கோடையில் கடுமையாகப் பிரகாசிக்கும்* என்ற வாக்கியத்தைப் பின்வருமாறு ஆயலாம்.



பொருளையும் தொடரியலையும் உள்ளடக்கிய காரணத்தால் தலை அடை அணுகுமுறை மொழியியலில் முக்கிய இடம் வகிக்கின்றது. ஒரு பொருள் (சாதனம்) எப்படி இருக்கிறது என்று கூறும் பெயரடை, பெயருக்குக் கூடுதலான விவரணையைத்/விளக்கத்தைத் தரும். வினையை

விவரிக்கும் வினையடை முறை, காலம், இடம் மற்றும் பிற செய்திகளைத் தரும். இவ்வகையிலான சொல் அமைப்பிற்கும் கருத்துரு அமைப்பிற்கும் இடையிலான பொருத்தம் மொழியைப் பொருளை வெளிப்படுத்துகிற ஒருமித்த ஒழுங்குமுறையாகப் பார்ப்பதால் பயனுள்ளதாய் இருக்கின்றது. தலை-அடை இலக்கணம் சார்பு இலக்கணமாக வடிவமைக்கப்பட்டு பயன்படுத்தப்படுகின்றது. சார்பு இலக்கணக் கொள்கை எவ்வாறு அமைப்பொழுங்கின் நிலையான அறிவு அமைப்பாக்கம் செய்யப்படுகின்றது. அல்லது வாக்கியம் ஆய்வு செய்யப்படுகின்றது என்று கூறாமல் அமைப்புகளை விளக்குவதின் வழிகளை வலியுறுத்துகின்றன.

5.2.1.1.அண்மை உறுப்பு அணுகுமுறை தொடரமைப்புச் சட்டகம் (Phrase Structure Formwork)

தொடரமைப்புச் சட்டகம் அண்மை உறுப்பு ஆய்வு அடிப்படையிலானது. அண்மை உறுப்பாய்வு ஒரு வாக்கியத்தை அண்மை உறுப்புகளாகப் பகுக்கும்.

அண்மை உறுப்பாய்வின் மாதிரிகள் (Models of IC analysis)

அண்மை உறுப்பாய்வு தொடரியல், உருபனியல், ஒலியனியல் என்ற எல்லா நிலைகளிலும் செய்யப்படுகின்றது. அண்மை உறுப்பாய்வில் பல மாதிரிகள் உள்ளன. அவற்றில் மூன்று வகை மாதிரிகள் கீழே தரப்பட்டுள்ளன.

1. அடையாளப்படுத்தாதப் பெட்டிப்படம்

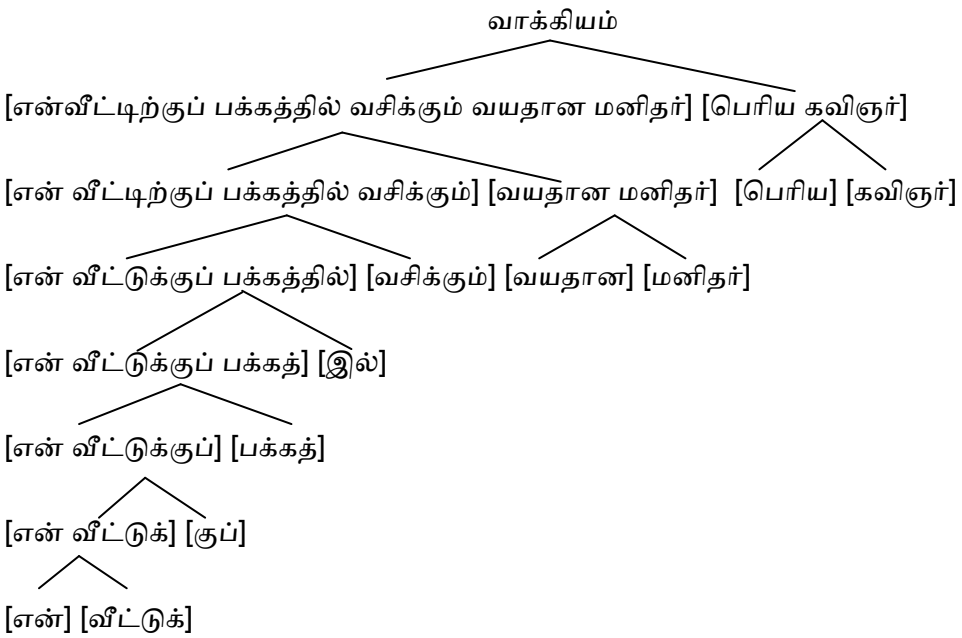
பின்வரும் எடுத்துக்காட்டில் தரப்பட்ட வாக்கியத்தின் அண்மை உறுப்புகள் பகுக்கப்பட்டு அடையாளப் படுத்தாமல் பெட்டிப்படமாகத் தரப்பட்டுள்ளது:

படம் 6

<table>
<tr><td>என்</td><td>வீட்டுக்</td><td>குப்;</td><td>பக்கத்</td><td>இல்</td><td>வசிக்கும்</td><td>வயதான</td><td>மனிதர்</td><td>பெரிய</td><td>கவிஞர்</td></tr>
<tr><td colspan="2">என் வீட்டுக்</td><td rowspan="2"></td><td rowspan="3"></td><td rowspan="4"></td><td rowspan="5"></td><td colspan="2" rowspan="5">வயதான மனிதர்</td><td colspan="2" rowspan="5">பெரிய கவிஞர்</td></tr>
<tr><td colspan="3">என் வீட்டுக்குப்</td></tr>
<tr><td colspan="4">என் வீட்டுக்குப் பக்கத்</td></tr>
<tr><td colspan="5">என் வீட்டுக்குப் பக்கத்தில்</td></tr>
<tr><td colspan="6">என் வீட்டுக்குப் பக்கத்தில் வசிக்கும்</td></tr>
<tr><td colspan="8">என் வீட்டுக்குப் பக்கத்தில் வசிக்கும் வயதான மனிதர்</td><td colspan="2" rowspan="2"></td></tr>
<tr><td colspan="10">என் வீட்டுக்குப் பக்கத்தில் வசிக்கும் வயதான மனிதர் பெரிய கவிஞர்</td></tr>
</table>

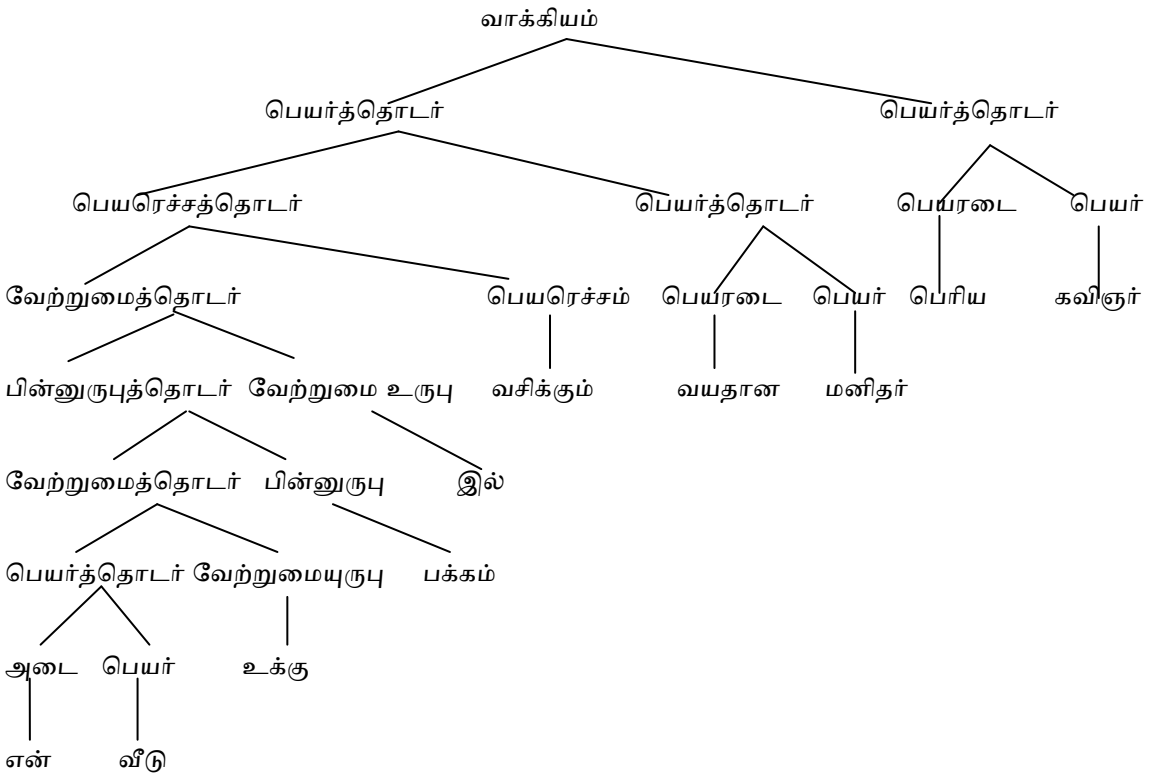
1. அடைப்புக்குறிப் படம்
2. [[[[என்] [வீட்டிற்]] [குப்]] [பக்கத்]] [இல்]] [வசிக்கும்]] [[வயதான] [மனிதர்]]] [[பெரிய] [கவிஞர்]]
3. கிளைப்படம்

வாக்கியப் பகுத்தாய்வைப் பின்வருமாறு படிநிலை அமைப்பில் அண்மை உறுப்புகள் கட்டுக்களை ஒன்றுக்குள் ஒன்றாக அமைவதைப் பின்வருமாறு கிளைப்படத்தில் காட்டலாடம்.



4. அடையாளப்படுத்தப்பட்ட படம்

இறுதி நிலையில் வரும் அண்மை உறுப்புகளைச் சொல்வகைப்பாட்டிற்காகவும் அவற்றால் விளையும்; கட்டுகளைத் தொடர்களாகக் கொண்டும் குறிப்பீடு செய்தால் பின்வருமாறு அடையாளப்படுத்தப்பட்ட படம் விளையும்.



அண்மை உறுப்புப் பகுப்பாய்வு குறித்த சில செய்திகள் கீழே பட்டியலிடப்பட்டுள்ளன.

1. அண்மை உறுப்பு பகுநிலை (analytic) அடிப்படையிலானது.

2. அடிப்படை மொழி அலகுகளைக் காண்பதற்காக அறிவியல் கண்டுபிடிப்புச் செயல்முறை (discovery procedures) செயல்பாட்டால் உருவானது.

3. வாக்கியங்களை இரண்டு அண்மை உறுப்புப் பாகங்களாகவும் ஒவ்வொரு சிறு பாகத்தையும் மேலும் சிறு அலகுகளாகவும் (உருபன் வகை) பகுப்பது இக்கோட்பாட்டின் கொள்கையாகும்.

4. சில மொழியியலார் வாக்கியத்தை பெரிய அலகுகளாகக் கொண்டு மேலிருந்து கீழ் நோக்கி (top-down) ஆய்வாளர்கள் சிலர் கீழிருந்து மேல்நோக்கி (bottom-up)ஆய்வார்கள்.

5. அண்மை உறுப்பு ஆய்வு அடையாளப்படுத்தப்பட்டால் மரபிலக்கணப் பகுப்பாய்வு (Parsing) போலவே அமையும். அதாவது வாக்கியத்தை இலக்கண வகைப்படுகளாக பகுத்தல் நிகழும்.

6. மரபு அடிப்படையிலான வாக்கியப் பகுப்பாயவும் தொடரமைப்பு இலக்கணம் அடிப்படையிலானதாகும். இவ்விலக்கணம் பின்வரும் கேள்விகளுக்கு விடை தருகின்றது.

1. ஒரு வாக்கியத்;தின் உருப்புகள் யாவை?

2. அவை எவ்வாறு ஒழுங்குபடுத்தப்பட்டுள்ளன

7. மரபிலக்கணங்கள் வாக்கியங்களையும் தொடர்களையும் விளக்கப் பின்வரும் கூற்றுகளைத் தருகின்றன.

1. ஒரு வாக்கியத்தில் பெயர்த்தொடரும் (Noun Phrase) வினைத்தொடரும் அடங்கும். 2. ஒரு பெயர்த்தொடரில் ஒரு பெயர் மட்டுமோ அடைகொளி அடை + பெயரோ அடைகொளி அடை + பெயரடை + பெயரோ வரும்.

எடுத்துக்காட்டாக,

1. அந்தப் பையன்

2. அந்த நல்லப் பையன்.

3. ஒரு வினைத்தொடரில் பெயர் மட்டுமோ வினை மட்டுமோ வினை + துணை வினையோ பின்வுருபுத் தொடர் + வேற்றுமைத் தொடர் + வினையோ வரும். எடுத்துக்காட்டு,

வந்தான்

வந்துக்கொண்டிருந்தான்

வீட்டிலிருந்து வந்துக்கொண்டிருந்தான் பழத்தைச் சாப்பிட்டான்

4. பின்னுருபுத் தொடரில் வேற்றுமைத் தொடரும் பின்னுருபும் வரும்.

எடுத்துக்காட்டு, வீட்டிலிருந்து

5. வேற்றுமைத் தொடரில் பெயர்த்தொடரும் வேற்றுமை உருபும் வரும்.

எடுத்துக்காட்டு,

வீட்டை, வீட்டின்

திட்டவட்டமற்ற சிக்கனமில்லாத இம்மாதிரியான விளக்கக்கூறுகள் தருவதற்குப்பதிலாக சில மரபு மற்றும் கருத்துச் சாயல் அடிப்படையில் ஒரு குழு ஆக்கவிதிகள் (Generation rules) அல்லது விரித்தெழுது விதிகள் (rewrite rules) தரலாம். அம்மாதிரியான விதிகளைத் தொடரமைப்பு விதிகள் (phrase structure rules) எனலாம். தொடரமைப்பு விதிகள் பின் மாதிரியான பொது விதிச் சுருக்கத்தைக் (rules-schema) கொண்டிருக்கும்.

அ → ஆ

அதாவது 'அ' என்ற குறியீட்டைத் தொடக்ககுறியீடு என்ற குறியீடாக விரித்தெழுதலாம். விரித்தெழுதும் செயல்பாட்டைத் தொடங்கிவைப்பதற்குப் பயன்படுத்தும் குறியீட்டைத் தொடக்ககுறியீடு எனலாம்.

எ.கா

வா → பெ.தொ + வி.தொ.

என்ற விரித்தெழுது விதியில் 'வா' என்பது தொடக்கக்குறியீடு. மேற்சொன்ன விதி ஒரு வாக்கியத்தில் பெயர்த்தொடரும் வினைத்தொடரும் வரும் என்று தெரிவிக்கின்றது. '+' என்ற குறி பிணைப்பைக் (concatenation) குறிப்பிடப் பயன்படுகின்றது. பெ.தொ, வி.தொ என்ற புலக்குறிகளை (labels) விரித்தெழுதலாம்.

எ.கா.

பெ.தொ → அடைகொளி அடை + பெயர்

என்ற விதி பெ.தொ. கணுவை விரித்தெழுதுகின்றது. (அடைகொளி அடையில் *அந்த, இந்த, என், உன், அவனது, அவளது, அதன், அவற்றின்* என்பன அடங்கும்). *பையன், சிறுமி* போன்றன பெயர்ச்சொல்லில் அடங்கும். கணு என்பது கிளைப்படத்தில் ஒன்றோ அல்லது அதற்கு மேற்பட்டகிளைகளோ வெளிவரும் தொடக்கவிடமாகும். தமிழில் வி.தொ என்பதை பி.உ.தொ + வே.தொ + வி.தொ + து.வி என விரித்து எழுதலாம்.

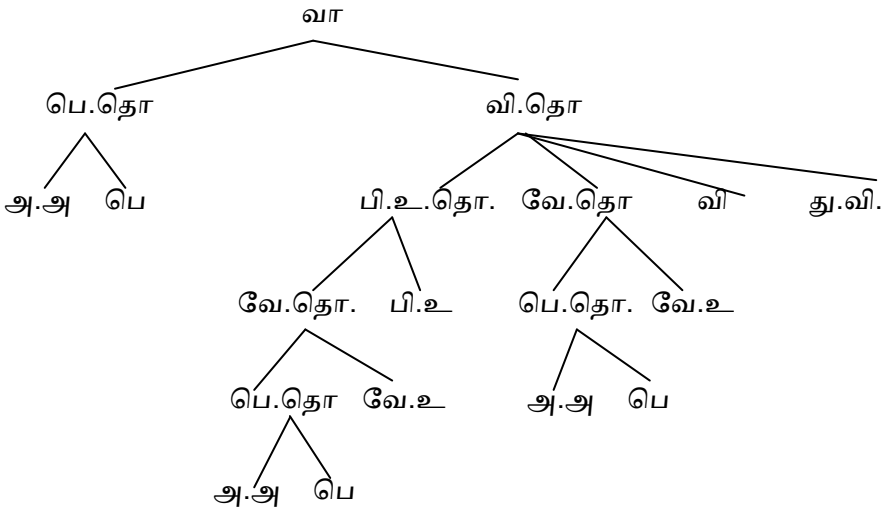
வி.தொ → பி.உ.தொ + வே.தொ + வி + து.வி.

பி.உ.தொ என்ற கணுவைப் பின்வருமாறு விரித்து எழுதலாம்.

பி.உ.தொ → பெ.தொ + பி.உ.

(வா = வாக்கியம், பெதொ = பெயர்த்தொடர், விதொ = வினைத் தொடர், பிஉதொ = பின்னுருபுத்தொடர், வேதொ = வேற்றுமைத் தொடர்; துவி = துணைவினை, வேஉ = வேற்றுமை உருபு, பிஉ = பின்னுருபு, அ அஅ = அடைகொளி அடை, விஅ =வினை அடை, பெஅ = பெயரடை, எஅ = எண்ணடை, மிஅ =மிகை அடை)

இதுவரை தந்த விதிகளைப் பின்வருமாறு கிளைபடமாகத் தரலாம்.



மேற்கண்ட கருத்துச் சாயல்களைப் பயன்படுத்தி தமிழ் இலக்கணத்தில் ஒரு பகுதிக்கு தொடரமைப்பு விதிகள் உருவாக்கலாம்.

வா →பெ.தொ + வி.தொ.

வி.தொ → வே.தொ + (பி.உ.தொ + வி +து.வி

து.வி → கொண்டிரு, இரு, வேண்டு, பார், போ

வி → கொல், வெட்டு, நட, பார், ஒடி

பெ.தொ → அ.அ + (பெ.அ) + பெ

பெ → பையன், பாம்பு, மரம், மேஜை, சிறுமி, பெண்.

அ.அ → அந்த, ஒரு, ஓர், தன்

பெ.அ → வயதான, இளமையான, அழகான

பி.உ.தொ → வே.தொ + பி.உ

வே.தொ → பெ.தொ + வே.உ

பி.உ → பக்கத்தில், மேலே, கீழே, நோக்கி, அருகில், அடியில்

வே.உ → ஐ, இல், .ஆல், உக்கு, ஓடு

மேற்கண்ட தொடரமைப்பு இலக்கணம் (Phrase sturcture grammar) பின் வரும் தமிழ் வாக்கியங்களை உருவாக்கும்.

அந்தப் பையன் ஒரு பாம்பைப் கொல்லப் பார்த்தான் (பெ.தொ + வே.தொ + வி + து.வி)

அந்த அழகான பெண் தன் வீட்டை நோக்கி நடந்தாள் (பெ.தொ + பி.உ.தொ + வி)

அந்த வயதான மனிதர் மரத்தை வெட்டப்போனார். (பெ.தொ + வே.தொ + வி + து.வி)

ஆனால் மேற்சொன்ன விதிகள் பின்வரும் பொருளற்ற வாக்கியங்களையும் உருவாக்கும்.

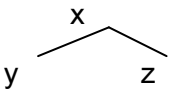
* மேஜை அந்தப் பெண்ணை மரத்திற்கு அருகில் கொன்றது.

* மரம் அந்தப் பையனைப் போகும்

* அந்தச் சிறுவன் அந்தச் சிறுமியை ஒடிக்க முடியும்

* மேஜை அந்த வயதான மனிதரைப் பார்க்கும்

மேற்சொன்ன விதிகள் சுழல் வரையறையில்லாத விதிகளாகும். அவற்றைச் சுழல் கட்டுண்ட விதிகளாக (Contex sensitive rules) மாற்றினால் மேற்சொன்ன வாக்கியங்களின் உருவாக்கத்தைத் தடை செய்யலாம். சூழல் வலையரை இல்லாத இலக்கணத்தைப் பயன்படுத்தி வாக்கியத்தைப் பகுத்தல், பகுப்பாய்வு/பகுத்தாய்தல் (parsing) எனப்படும். உள்ளீடு செய்யப்படும் கோர்வையைத் (input string) தெரிந்து கொள்வதும் அதற்கு அமைப்பைத் தருகிறதும் ஆன இரண்டும் இணைந்த செயல்பாடுதான் பகுத்தாய்தலாகும். வாக்கியத்தை தெரிந்து கொள்வதும் அதற்கு அமைப்பைத் தருகிறதும் ஆன செயல்பாடு தொடரியல் பகுத்தாய்தல் (syntactic parsing) எனப்படும். பகுத்தாயும் கருவி பகுத்தாய்வான் (parser) எனப்படும். இவ்வாறு பகுத்தாய்வான் இரு செயல்களைச் செய்ய வேண்டும். ஒன்று ஒரு கோர்வைத் தரப்படும் போது பகுத்தாய்வான் முதலில் அதைப் பகுத்தாய வேண்டிய மொழியின் வாக்கியம் என்பதைத் தெரிந்து கொள்ள வேண்டும். இந்நிலையில் பகுத்தாய்வானுக்கு உள்ளடங்கிய தெரிந்துகொள்வான்கள் (built-in recognizers) இருக்க வேண்டும். இரண்டாவது அது வெளியீடு செய்கிற அமைப்பை வாக்கியத்திற்கு தர வேண்டும். இலக்கணத்தில் இருக்கிற மொழியியல் செய்திகளை பகுத்தாய்வான்கள் சார்ந்திருக்க வேண்டும் என்பது இதன் மூலம் தெரிகின்றது. வெளியீட்டு வசதிகளுடன் கூடிய தெரிந்துகொள்வான்களைப் (recognizers) பொறுத்த வரையில் இலக்கணங்களும் பகுத்தாய்வான்களும் பகுத்தாயும் பண்பு வாய்ந்தவை. x→y என்ற விதி ஒரு உருவாக்கமாகத் தரப்பட்ட கலவைநிலை கட்டளையாகக் கருதப்பட வேண்டும்.



பகுத்தாய்வான் பெரும்பாலும் குறியீடுகளின் கோர்வையை ஏற்கும்: அதற்கு ஒரு விதிகையைப் பயன்படுத்தும் அது ஒரு கோர்வையின் விரித்தெழுது விதி போல் வரும். எடுத்துக்காட்டாக b → d என்ற விதியைப் பயன்படுத்தி abc என்பதை adc என்று விரித்தெழுதலாம்; adc என்பதை adc என்றும் விரித்தெழுதலாம். 'ABC' 'ADC', மற்றும் 'Adc' என்பன வாக்கிய வடிவுகள் (Sententical forms) அல்லது ஆக்கங்கள் எனப்படும். 'ADC' என்ற கோர்வை 'ABC'-யில் இருந்து ஒரு விதியைப் பயன்படுத்தி பெறப்படுவதால் ADC-ஆனது ABC-யிலிருந்து நேரடியாகப் பெறப்பட்டது ஆகும். 'ADC' என்ற கோர்வை 'ABC'-யில் இருந்து ஒரு விதியைப் பயன்படுத்தி பெறப்படுவதால் ADC-ஆனது ABC-யிலிருந்து மறைமுகமாகப் பெறப்பட்டது ஆகும். ஒவ்வொரு படியிலும் பகுத்தாய்வான்கள் ஏதோ ஒரு அமைப்பை வெளியிடும். ஒரு குழும விதிகளின்படி வரக்கூடிய எல்லா அமைப்புகளையும் நாம் தெரிந்தால் ஒருவாக்கியத்தைப் பகுத்தாயலாம்.

5.2.1.2. பகுத்தாயும் நடைமுறைத்திறன் (parsing strategies)

மொழியியல் அறிவைப் பிரதிபலிக்கின்ற விதிகளின் அடிப்படையில் பகுத்தாய்வான் செயல்படும். வாக்கியங்களுக்கு அமைப்புத் தருகையில் ஒரு குழும விதிகளைப் பல நிரல்களில் நிறைவேற்றலாம். ஒவ்வொரு வேறுபட்ட நிரலும் வேறுபட்ட பகுத்தாய்தல் நடைமுறைத் திறத்துடன் தொடர்புடையது. பகுத்தாய்வான்களை அவற்றோடு தொடர்புடைய நடைமுறைத்திறன் அடிப்படையில் வகைப்பாடு செய்யலாம். பகுத்தாயும் நடைமுறைத் திறனாக இரண்டு அளவீடுகள் அடிக்கடி பயன்படுத்தப்படுகின்றனவும் நிலைபேறு பெற்றவனவும் ஆகும். முதலாவது அது பகுத்தாயும் கோர்வைக்கு வெளியீடுகளின் மொழியியல் அமைப்பைக் கவனக் குவிப்பு செய்கிறது. உள்ளீட்டுக் கோர்வையிலிருந்து தொடங்கி அமைப்பு கட்டப்படுவதனால் அந்த பகுத்தாய்வான் கீழிருந்து மேல் (bottom-up) செயல்படுவதாகக் கூறலாம். அல்லது தொடக்கக் குறியீட்டிலிருந்து (கிளை அமைப்பின் வேர்) தொடங்குவதனால் பகுத்தாயவான் மேலிருந்து கீழ் (top-town) செயல்படுவதாகக் கூறலாம். பகுத்தாயும் நடைமுறைத் திறன்களை வகைப்படுத்தும் அளவீடுகளை ஒரு எடுத்துக்காட்டு மூலம் விளக்கலாம். பின்வரும் ஒரு குழும விரித்தெழுது விதிகளை எடுத்துக்கொள்வோம்.

1a. s→ AB

1b. s→ cd

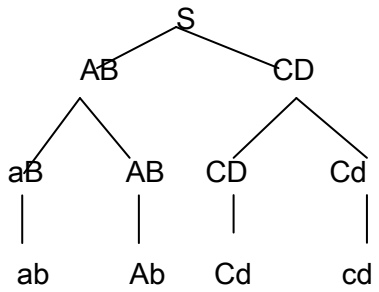
1c. A→ a

1d. B→ b

1e. C→ C

1f. D→d

தரப்பட்டுள்ள ஒரு குழும விதிகளிலிருந்து அவற்றால் பெறப்படும் எல்லா ஆக்கங்களின் திட்டத்தைக் கட்ட முடியும். தொடக்க குறியீடான S-இல் இருந்து இரண்டு விதிகளைச் செயல்படுத்தலாம். விதி 1a AB-இன் ஆக்கத்தையும் விதி 1b CD-இன் ஆக்கத்தையும் விளைவிக்கும். AB-யில் விதி 1c -ஐயும் விதி 1d-ஐயும் பயன்படுத்தி முறையே aB மற்றும் Ab என்ற கோர்வைகளைப் பெறலாம். இந்த எண்ணத்தைப் பின்பற்றி தரப்பட்ட ஒரு குழும விரித்தெழுது விதிகள் அனுமதிக்கும் ஆக்கங்கங்களின் எல்லா கோர்வைகளையும் கண்டுகொள்ளலாம். விளைவு பெரும்பாலும் கிளை வடிவாக உருப்படுத்தம் செய்யப்படும்.



மேலே தரப்பட்ட கிளை அமைப்பு 1 இல் தரப்பட்ட விதிகளால் விளையும் ஆக்கங்களின் எல்லா கோர்வைகளையும் காட்டுகின்றது. கிளையில் வரும் ஒவ்வொரு கணுவும் (AB,CD, aB, ab, Ab, CD, Cd, cd) வாக்கிய வடிவில் நடைமுறைப்படுத்தக் கூடிய வேறுபட்ட விதிகளின் தேர்வை உருப்படுத்தம் செய்யும். கிளை அமைப்பின் கிளைகள் (ab, cd) அதாவது அடியில் வரும் கணுக்கள் மேலும் விதிகளின் பயன்படுத்த இயலாத வாக்கிய வடிவுகளை உருப்படுத்தம் செய்யும். இந்த கிளை அமைப்பை மொழியியலார் எவ்வாறு ஒரு வாக்கியத்தின் பாகங்கள் இணைகின்றன என்பதைக் காட்டுகின்ற மொழியியல் அமைப்பை உருப்படுத்தம் செய்யப் பயன்படுத்துகின்ற கிளையுடன் ஒப்பிட்டு குழப்பிக்கொள்ளக்கூடாது. மேல் சொன்ன விதிகள், கிளை அமைப்பைக் காட்ட இயலும் ஆக்கங்களின் எல்லாக் கோர்வைகளையும் தருகின்றன. இந்த வகையான கிளை அமைப்பின் அடிப்படையில் ஆழம் முதல் (depth-first) அல்லது அகலம் முதல் (breadth- first) என்ற பகுத்தாயும் நடைமுறைத்திறன்களின் பாகுபாடு அமையும்.

5.2.1.2.1. மேலிருந்து கீழ் மற்றும் கீழிருந்து மேல் பகுத்தாய்தல் (Top- Down) Versus bottom-up Parsing)

பின்வரும் இரு எடுத்துக்காட்டுகள் எவ்வாறு மேலிருந்து கீழ் மற்றும் கீழிருந்து மேல் பகுத்தாய்வான்கள் பின்வரும் வாக்கியத்திற்கு அமைப்பு தருகின்றது என்பதை விளக்கும்.

1. ஒரு பூனை எலியைப் பிடித்தது.

இரண்டும் பின்வரும் விரித்தெழுதும் விதிகளைப் பயன் படுத்தப்படுவதாக வைத்துக்கொள்வோம்.

2.a. வா →பெ.தொ + வி.தொ

2. b. பெ.தொ →அ.அ + பெ

2. c. வி.தொ →வே.தொ + வி

2. d. வி.தொ →வி

2. e. வே.தொ →பெ.தொ + வே உ

2. f. வி →பிடித்தது

2. g. பெ→ பூனை

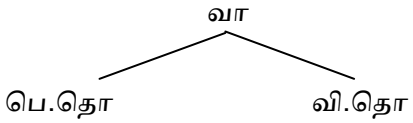
2. h. பெ →எலி

2. i. அ.அ →ஒரு

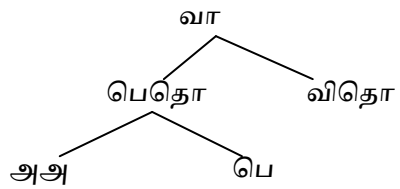
2. j. வே.உ→ ஐ

5.2.1.2.2. மேலிருந்து கீழ் பகுத்தாய்தல் (மேல்-கீழ் பகுத்தாய்த்தல்) (Top- Down parsing)

மேலிருந்து கீழ் பகுத்தாய்வான்கள் எப்பெழுதும் தொடக்கக் குறியீடான 'வா' என்பதிலிருந்து தொடங்கும் அதில் பயன்படுத்த விதிகளைக் கண்டு கொள்ளும் அதை விரிக்கும் எடுத்துக்காட்டில் 2.a என்ற விதிதான் இருக்கிறது. அதைப்பயன்படுத்தி விளையும் விளைவு பின்வரும் அமைப்பைத் தரும்.



இரு புதிய கணுக்கள் தோன்றும். இந்த கணுக்கள் இறுதிக் குறியீடுகளா (terminal symbols) என்று முதலில் பகுத்தாய்வான் பார்க்கும். அப்படியானால் அந்தக் குறியீடுகள் பகுத்தாயப்படுகின்ற கோர்வையுடன் பொருத்திப்பார்க்கப்படும். இல்லாவிடில் பகுத்தாய்வான் முதல் இறுதியுறா குறியீட்டுக் கணுவை விரிக்கும். மேற்கண்ட எடுத்துக்காட்டில் பெதொ வில் 2.b என்ற விதி பயன்படுத்தப்பட்டு பின்வரும் அமைப்பு விளையும்.



புதிதாக உருவாக்கப்பட்ட இந்தக் கணுக்களும் இறுதிக் கணுக்கள் அல்ல.

கீழிருந்து மேல் பகுத்தாய்தல் (Bottam-up parsing)

கீழ்-மேல் பகுத்தாய்தல் உள்ளீடு செய்யப்பட்ட கோர்வையைத் தொடக்கமாகக் கொண்டு செயல்படத் தொடங்கி அதை வேரான 'வா' என்று சுருக்கும். அது ஒரு வாக்கியத்தை உள்ளீடாக ஏற்கும் சொற்களை (இறுதியுற்ற குறியிடுகளை) அவற்றின் வகைப்பாட்டால் இடம்பெயர்க்கும். இதற்காக அது விரித்தெழுது விதிகளின் வலப்பக்கம் பார்த்து அவற்றை இடப்பக்கம் உள்ள வகைப்பாடாகச் சுருக்க வேண்டும். மீண்டும் *ஒரு பூனை எலியைப் பிடித்தது* என்ற வாக்கியம் 2a-j

என்ற விதிகளால் கூறப்பட்டுள்ள இலக்கணத்தால் அமைப்பு பெறும். எடுத்துக்காட்டாக முதலில் 2f. என்ற விதியைப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்.

வி
↓
ஒரு பூனை ஒரு எலி ஐ பிடித்தது

2.j.-ஐப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்

வே.உ வி
↓ ↓
ஒரு பூனை ஒரு எலி - ஐ பிடித்தது

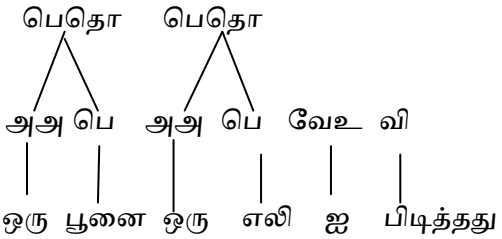
2.h.-ஐ பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்

பெ பெ வே.உ வி
↓ ↓ ↓ ↓
ஒரு பூனை ஒரு எலி ஐ பிடித்தது

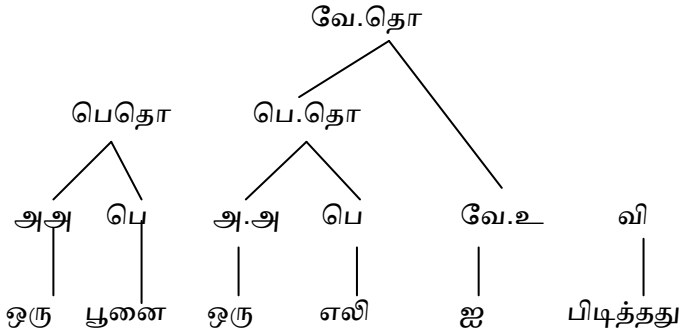
2 .i.-ஐப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்

அஅ பெ அஅ பெ வே.உ வி
↓ ↓ ↓ ↓ ↓ ↓
ஒரு பூனை ஒரு எலி ஐ பிடித்தது

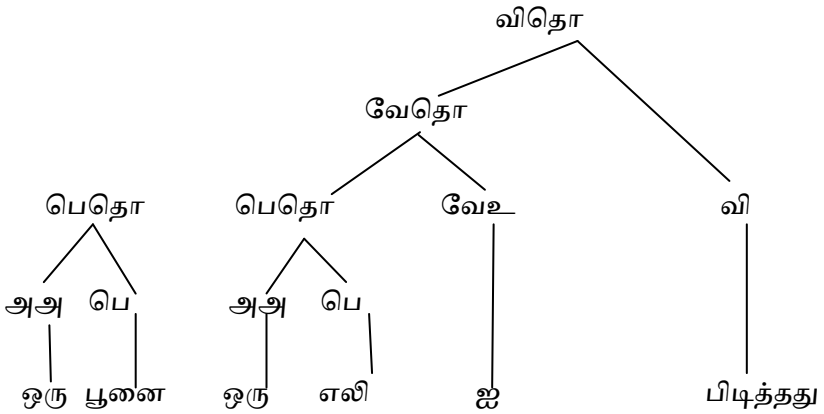
இந்த இடத்தில் 2b என்ற விதி ‘அ.அ’ என்பதையும் ‘பெ’ என்பதையும் இணைந்து அவற்றை பெ.தொ எனக் குறிக்கும்.



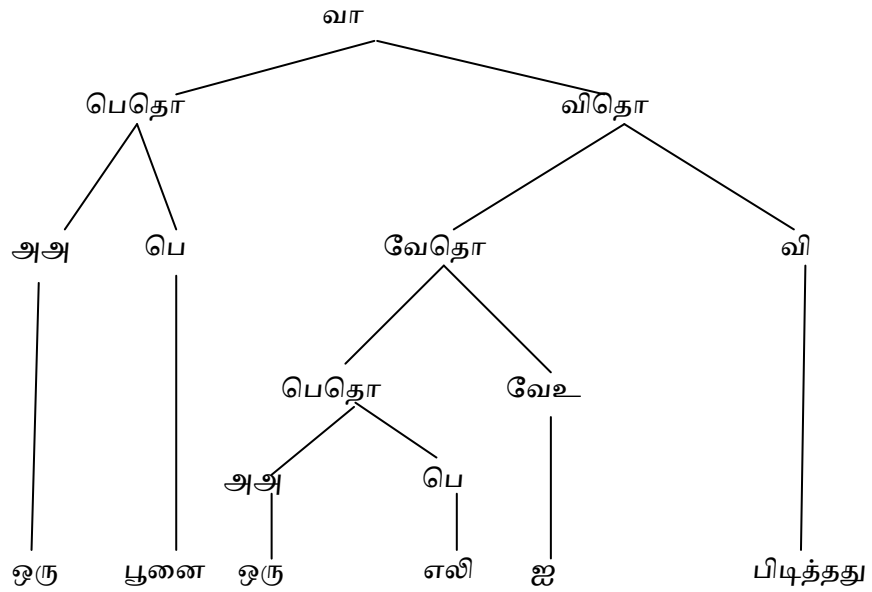
இந்த இடத்தில் 2e என்ற விதி வேஉ என்பதையும் பெதொ என்பதையும் இணைத்து அவற்றை வேதொ என சுருக்கும்.



விதி 2c ஐ-வேதொ ஐயும் வி-ஐயும் இணைத்து வி.தொ. எனச் சுருக்கும்.



விதி 2a பெ.தொ ஐயும் வி.தொ ஐயும் இணைத்து வா ஆகச் சுருக்கும்.



பகுத்தாய்வான் வாக்கியத்தின் வலப்பக்கத்திலிருந்து தொடங்கி கோவையின் முன்னோக்கிச் செயல்படுவதாகக் கொள்வது அவ்வளவு சரியானதல்ல. தெளிவாகக் கூறினால் மொழி இம்மாதிரி செயல்படுவதில்லை. இதன் அடிப்படையில் கீழ்-மேல் பகுத்தாய்வான் உளவியல் அடிப்படையில் சரியானது அல்ல என்று வாதிடலாம். வேறுபட்ட ஆக்கங்களின் தொடர்ச்சியைப் பின்பற்றினாலும் ஒரே இலக்கணத்தைப் பின்பற்றுவதனால் இடது பக்கத்திலிருந்து தொடங்கும் பகுத்தாய்வானும் வலது பக்கத்திலிருந்து தொடங்கும் பகுத்தாய்வானும் பகுத்தாய்தல் அடிப்படையிலும் பெற்ற விளைவின் அடிப்படையிலும் ஒன்றாகும். இந்த வலமிருந்து இடம் என்ற வேறுபடலுக்கும் வலது பகுத்துத்தாய்தல் அல்லது இடது பகுத்தாய்தல் என்பதற்கும் தொடர்பில்லை. வலது பகுத்தாய்தலானது கீழ்-மேல் பகுத்தாய்தலின் விளைவாகும். இது விதிகளின் வலப்பக்கத்தில் காணப்படும் குறியீடுகளைப் பார்த்து வாக்கிய வடிவுகளைச் சுருக்கும். ஒரு மேல்-கீழ் பகுத்தாய்வான் இடது பகுத்தாய்தலை நடைமுறைப்படுத்தும். இது விதிகளின் இடப்பக்கத்தில் காணப்படும் குறியீடுகளை விரித்து வாக்கிய வடிவுகளை ஆக்கும். இதுபோல் ஒரே மொழியியல் செய்திகளின் படி ஒரே வாக்கியங்களுக்கு ஒரே அமைப்புகளைத் தருவதால்

ஒரே இலக்கணத்தை நோக்கீடு செய்யும் மேல்-கீழ் பகுப்பானும் கீழ்-மேல் பகுப்பானும் ஒன்றாகும். அவைகள் வேறுபட்ட ஆக்கங்களின் தொடர்ச்சியைப் பின்பற்றுகிறது என்ற உண்மையுடன் எவ்வளவு நினைவகம் தேவை எவ்வளவு கணிப்பு நேரத்தை உட்படுத்துகிறது என்பதாலும் வேறுபடும்.

5.2.1.2.2. ஆழம் முதல் மற்றும் அகலம் முதல் பகுத்தாய்தல் (Depth first Versus breath first Parsing)

பின்வரும் ஒரு குழும விரித்தெழுது விதிகளை எடுத்துக் கொள்வோம்.

3 a. S ⟶ AB

3 b. S ⟶ CD

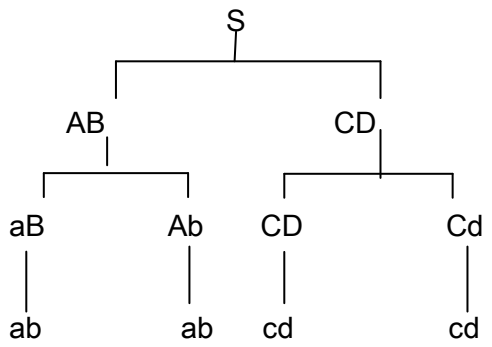
3 c. S ⟶ a

3 d. B ⟶ b

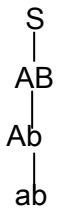
3e. C ⟶ c

3 f. D ⟶ d

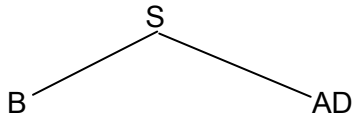
இவற்றோடு தொடர்புடைய பின்வரும் ஆக்கக் கிளையமைப்பையும் எடுத்துக் கொள்வோம்.

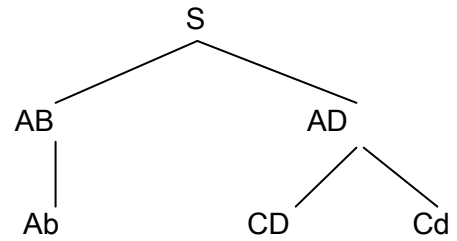


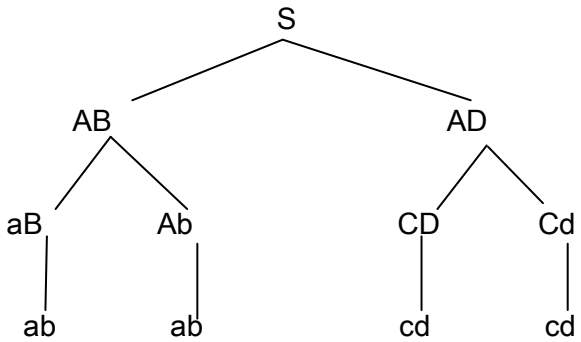
இந்த கிளை அமைப்பை இருவிதமாக அணுகலாம்: ஒன்று அதன் செங்குத்து நிலையில் கவனம் செலுத்திப் பெறலாம்; இரண்டு அதன் இடை நிலையில் கவனம் செலுத்திப் பெறலாம். இந்த இரு வேறுபட்ட நோக்கும் பகுத்தாயும் நடைமுறைத்திறனின் வேறுபாட்டை விளைவிக்கும். ஆழம் முதல் பகுத்துக்குறித்தல் நாம் மேலே தரப்பட்டுள்ள கிளை அமைப்பின் செங்குத்து வழியை எடுத்துக் கொள்வோம். வேரை வாக்கியத்துடன் இணைக்கும் ஒரு தனி செங்குத்து விதியைத் தேர்ந்தெடுப்போம். எடுத்துக்காட்டாக கிளைக் கணுவில் அடங்கும் ab என்ற வாக்கியத்திற்கு பின்வருமாறு வழியைத் தேர்ந்தெடுக்கலாம்.



இந்த வழி கிளை அமைப்பில் வரும் எந்தச் செங்குத்தான வழியைப் போல வாக்கிய வடிவுகளின் ஒரு தொடர்ச்சியைத் தரும். ஒவ்வொரு வாக்கிய வடிவும் வேருடனோ அல்லது வேறு தனி விதியால் விளைந்த வாக்கிய வடிவுடனோ ஒரு தனி விதியைப் பயன்படுத்துவதால் விளைவதாகும். ஒரு குறிப்பிட்ட ஆக்கத்திற்கு பல விதிகளைப் பயன்படுத்தலாம் என்றாலும் ஒரே ஒரு விரித்தெழுது விதிதான் தக்கவைக்கப்படும் பிற விருப்புகள் கிளையின் பிற வழிகளில் வெளிப்படுத்தப்படும். ஒரு தனிச் செங்குத்தான வழியாகப் பின்பற்ற இயலாது. ஒரு கிளை அமைப்பில் ஒரு தனி செங்குத்தான வழியிலே உருப்படுத்தம் செய்யத்தக்க வாக்கிய வடிவுகளைப் பின்பற்றும் பகுத்தாய்வான் ஆழம்-முதல் பகுத்தாய்வான் என்று அழைக்கப்பெறும். மேல்-கீழ் மற்றும் கீழ்-மேல் பகுத்தாய்வான்கள் இவ்வகையைச் சாரும். மேற்கண்ட கிளைப்படம் ஒரே வாக்கியத்தை ஒன்றுக்கும் மேற்பட்ட வழிகளில் தொடக்கக் குறியீடான 'வா' என்பதுடன் இணைக்கலாம் எனக் காட்டும். ஆழம் முதல் பகுத்தாய்வான் ஒரு சமயத்தில் ஒரு வழியை ஆய்வதால் தொடக்கத்தில் தேர்ந்தெடுக்கப்பட்ட பாதை தவறாகப் போகலாம். இந்த இடங்களில் பகுத்தாய்வான் சரிசெய்து தவறிலிருந்து விடுபட வேண்டியது அவசியம் இதன் காரணமாக ஆழம் முதல் பகுப்பான்கள் பின்னழுவல் வசதியுடன் நிறைவேற்றப்பட வேண்டும். ஆழம் முதல் பகுத்தாய்வான் நாம் ஒரு ஆக்கக்கிளை அமைப்பை இடைநிலைப் பரிமாணத்திற்கு முக்கியத்துவம் கொடுத்து கிளை அமைப்பின் ஒரு நிலையில் எல்லாக் கணுக்களையும் எடுத்துக் கொண்டு பார்க்கலாம். எடுத்துக்காட்டாக ஆக்கக்கிளை அமைப்பின் வேர் 'S' என்ற வாக்கிய வடிவைத் தாங்கும் கணுவைக் கொண்டிருக்கும். AB மற்றும் AD என்ற வாக்கிய வடிவுகளைக் கொண்ட இரண்டு மகள் கணுக்கள் 'S' என்பதிலிருந்து வேறுபட்ட விதிகளால் விளைந்தனவாகும். இம்மாதிரியான நிகழ்வில் பகுத்தாய்வான் ஒரே நேரத்தில் வேறுபட்ட ஆக்கங்களை உருவாக்கும். அடுத்த அடியில் இரண்டின் விளைவுகளின் மேல் இயலும் எல்லா விதிகளும் செயல்படுத்தப்படும். இவ்வகையிலான பகுத்தாய்வான் அகலம் முதல் பகுத்தாய்வான் எனப்படும்.







வீதி முதல் பகுத்தாய்வான்கள் உருவாக்கப்பட்ட எல்லா வாக்கிய வடிவிலும் இயலும் எல்லா விதிகளைச் செயல்படுத்தும் அவை ஆக்கக் கிளை அமைப்பின் இடைநிலைப் பரிமாணத்தை ஆயும். ஒரே நேரத்தில் வரும் எல்லாத் தேர்வுகளையும் தீர்த்து விட்டு தோல்வி அல்லது வெற்றி என்ற முடிவுக்கு எடுத்துச் செல்லும். இவ்வகையிலான செயல்முறை தோல்வியின் போது பின்னழுவலை அதிக பட்சமாகும். தவறான தேர்வால் விளையும் ஆக்கத் தொடர்ச்சி இல்லாது போனாலும் ஒரே சமயத்தில் உருவான எல்லா வெற்றிப் பிறழ்ச்சிகளும் எஞ்சும்.

வகைப்பாடு (Classification)

	ஆழம் முதல்	அகலம் முதல்
மேல்-கீழ்	மேல்-கீழ் ஆழம்-முதல்;	மேல்-கீழ் அகலம்-முதல்
கீழ்-மேல்	கீழ்-மேல் ஆழம்-முதல்	கீழ்-மேல் அகலம்-முதல்

5.2.1.2.3.நிலைமாற்ற வலைப்பின்னல்களால் பகுத்தாய்தல் (Parsing with Transition Network)

நிலைமாற்றப் பின்னல் இலக்கணங்கள் கணினி மொழியியலில் மிகவும் வெற்றிகரமாக நடைமுறைப்படுத்தப்படுகின்றது. இதன் அடிப்படைக் கொள்கையில் அமைந்த பல வித்தியாசமான நிலைமாற்றபின்னல்கள் இருக்கின்றன. நிலைபேறு பெற்ற ஒன்று விரிவாக்கப்பட்ட நிலைமாற்றப் பின்னலாகும் (Augmented Transition Network (ATN)). இது உட்ஸ் (Woods, 1970) என்பவரின் மாதிரியாகும். இது ATN என சுருக்கி அழைக்கப்படும். இலக்கணங்களும் தானியங்கிகளும்: ஒரு குழும விதிகள் ஒரு மொழியின் வாக்கியங்களின் விளக்கமாக எடுத்துக்கொள்வதோடு அந்த மொழியின் கோர்வைகளை ஏற்கும் அல்லது உருவாக்கும். அருவப்பொறியின் (abstract machine) விளக்கமாகவும் எடுத்துக்கொள்ளலாம். இம்மாதிரியான பொறிகள் தானியங்கி என்று அழைக்கப்படும். இவற்றின் சாத்தியமான நடத்தை இவைகள் பொருத்தப்படுகின்ற இலக்கணத்தின் விதிகள் உட்கலவை நிலையில்

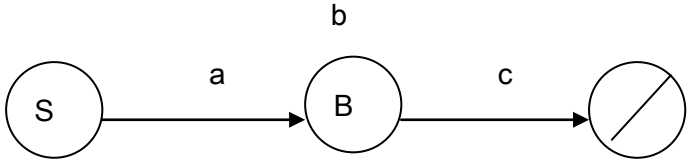
நிர்ணயிக்கப்படுகின்றது. இம்மாதிரியான எல்லாப் பொறிகளையும் ஒரு குழுமநிலைகளாலும் (set of states) ஒரு நேரத்தில் ஒரு குறியீட்டை எடுக்கும் உள்ளீட்டுத்திட்டத்தாலும் (input device) உள்ளீட்டைப் பரிசோதித்துப் படித்து பொறியை ஒரு நிலையிலிருந்து அடுத்த நிலைக்கு மாறச் செய்யும். ஒரு கட்டுப்பாட்டுத் தொகுதியாலும் (control unit) பண்பாக்கம் செய்யலாம். குறியீடுகளைச் சேகரிக்கவும் பரிசோதிக்கவும் இதைக் கட்டுப்பாட்டுத் தொகுதியில் அடைய இயலும். பொறியின் கணிப்புத்திறன் அதன் நினைவகத்தின் கூட்டுநிலையாலும் நினைவகத்தில் நிறைவேற்றக்கூடிய செயன்மைகளாலும் நிர்ணயிக்கப்படும். எளிமையான தானியங்கி முற்றுநிலைப் பொறிகள் (Finite State Machines (FSM) முற்றுநிலைத் தானியங்கிகள் (Finite state Automata) எனப்படும். அவைகள் ஏற்கும் மொழியைச் சீரான மொழிகள் என்பர். சீரானமொழிகளுக்கு சூழல் வரையறை இல்லாத மொழிகளின் (மத்திய, உட்படை/ உட்படுத்தம், தவிர்த்த) எல்லா பண்புகளும் இருக்கும். இதனால் இதை முற்று எண்ணிக்கை நிலைகள் (finite number of states) உள்ள தானியங்கியால் தெரிந்துகொள்ள இயலும். முற்றுநிலைப் பொறிகளுக்கு அக நினைவகம் (Internal memory) இல்லை. இதன் காரணமாய் அதன் நடத்தை அப்போது படிக்கப்படுகிற குறியீட்டாலும் அப்போதைய பொறி நிலையிலும் நிர்ணயிக்கப்படும். இம்மாதிரியான பொறியின் நடத்தையை எடுத்துக்காட்ட பின்வரும் எளிய இலக்கணத்தை எடுத்துக்கொள்வோம்.

S ⟶ aB

B ⟶ bB

B ⟶ c

இது ab*c என்ற சீரான மொழியை உருவாக்கும். இதற்கு சமமான தானியங்கியை மரபு அடிப்படையில் ஒரு நிலைப்படம் (State diagram) மூலம் உருப்படுத்தம் செய்யலாம்.



மேலே உள்ள படத்தில் கணுக்கள் நிலைகளைக் குறிப்பிடும். விற்கள் (படத்தில் அம்பு வடிவத்தில் தரப்பட்டுள்ளது) நிலைமாற்றங்களைக் குறிப்பிடும். விற்களில் உள்ள புலக்குறிப்புகள் (labels) ஒரு நிலைமாற்றம் சம்பவிப்பதற்கான கட்டுப்பாடுகளைத் தரும். குறுக்குகோடு அடையாளப்படுத்தப்பட்ட கணுவுக்கு ஒரு தனிப்பட்ட நிலை உண்டு. அது இறுதிநிலை எனப்படும். விரித்தொழுது விதிகளிலிருந்து நிலை வரைபடத்திற்கு பெயர்ப்பு செய்வது இந்த தானியங்கிக்கு நேரிடையானது. இறிதியுறா குறியீடுகள் (non-terminal symbols) நிலைகளுக்குச் (states) செல்லும். இறுதியுற்ற குறியீடுகள் (terminal Symbols) நிலை மாற்றத்திற்குச் (transition) செல்லும். இந்த தானியங்கியின் நடத்தையைப் பின்வருமாறு விளக்கலாம். கணித்தலானது

(computation) தொடக்கநிலை (initial satate) என்று குறிப்பிடப்பட்ட நிலையில் தொடங்குகிறது. (இங்கு S தொடக்கநிலையாகும்). உள்ளீடுகளில் உள்ள முதலாவது குறியீடு அதோடு பொருந்துகிற நிலைமாற்றம் இருக்கிறதா என்று பரிசோதனை செய்யும். அவ்வாறு பொருந்தினால் பொறி நிலை மாற்றத்தைத் தொடர்ந்து அடுத்த நிலைக்குச் செல்லும் மற்றும் உள்ளீட்டுக் குறியை ஏற்றுக்கொள்ளும். அதாவது அது தனது கவனத்தைத் தொடர்ச்சியிலுள்ள அழுத்தக் குறியீட்டிற்குத் திரும்பும். இந்தச் செயற்பாங்கு பொறி நிறுத்துவது வரை தொடர்ந்து நடக்கும். அதாவது கூடுதல் நிலைமாற்றங்கள் இல்லாதது வரை நடக்கும். பொறி இறுதிநிலையில் நிறுத்தும் போது நம்கோர்வையை ஏற்க்கப்பட்டது எனலாம். இதே தானியங்கியை உள்ளீட்டிலிருந்து வெளியீட்டு நிலைமாற்றச் செயலின் பண்பை மாற்றி கோர்வைகளை உருவாக்கச் செய்யலாம். நாம் இறுதி-உள்ளே-முதல்-வெளியே (Last-in-First-Out) பட்டியலுக்குக் கட்டுப்படுத்தப்பட்ட நினைவகத்தையும் (FSA Memory) அதை திறமையாகக் கையாள (manipulate) கட்டுப்பாடு மரபுகளையும் (control convention) சேர்த்தால் தானியங்கியைக் கீழே தள்ளு தானியங்கியாக (Push Down Automata) மாற்றலாம். அது கூடுதலான நினைவகத்தின் நன்மை காரணமாக மத்திய உட்பொதித்தலைக் (centre embedding) கையாளலாம். கீழே தள்ளு தானியங்கிகள் சூழல் வரையறையற்ற மொழிகளின் (Context Free Language(CFL) வகுப்புகளை ஏற்க இயலும்.

5.2.1.2.3.1. மறுதரவு நிலைமாற்ற வலைப்பின்னல்கள் (Recursive Transition Network (RTNS))

பெரிதாக்கப்பட்ட நிலைமாற்றப் பின்னல் (Augmended transition Network) என்பது நிலைமாற்ற மறுதரவு வலைப்பின்னல் (Transition recursive network) என்று அழைக்கப்படுகின்ற அபூர்வமான சூழல் வரையறையற்ற தெரிந்துகொள்ளும் திட்டத்தின் (Ingenious Context Free Recognition Device) விரிவாக்கமாகும். பெரிதாக்கப்பட்ட நிலைமாற்றப் பின்னல்கள் இலக்கணங்களின் விரித்தொழுது விதி உருப்படுத்தத்திலிருந்து விரும்பத்தக்க மாற்றமாக அமைகின்றது. இது மறுதரவு நிலைமாற்ற வலைப்பின்னல்களின் அடிப்படைக்காரணம் ஆகும். பின்வரும் எளிதான இலக்கணத்தை எடுத்துக்கொள்வோம்.

விதி 1. அ: வா → பெ.தொ + வே.தொ * + வி + (து.வி)

விதி 1. ஆ: வா → பெ.தொ + பெ.தொ + (து.வி)

விதி 2. பெ.தொ → (அ.அ) (அளவடை) பெ.அ * பெ*

விதி 3 : வே.தொ →பெ. தொ + வே.உ

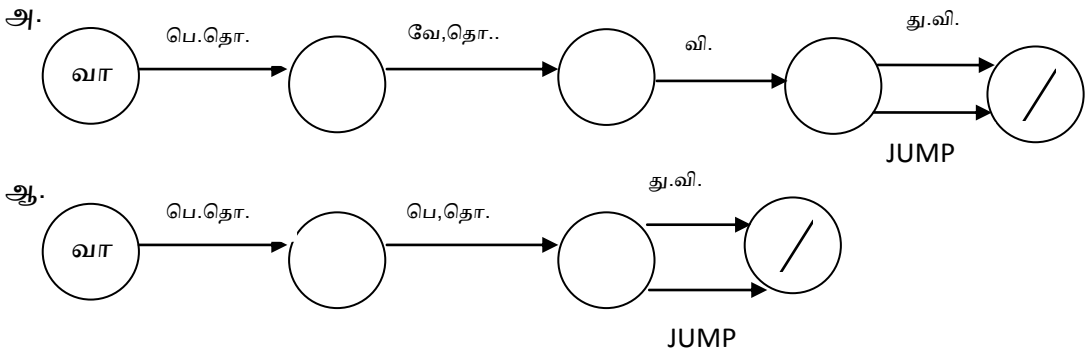
இவை சூழல் இல்லாத மொழியை {து.வி, வி, அ.அ, அளவடை, பெ.அ, வே.தொ.} என்ற எழுத்துக்கூட்டலால் விளக்குகின்றது. இந்த இலக்கணத்திற்கு குறைந்த ஆக்க சாத்தியம் இருக்கின்றது. பிறை அடைப்புக்குறிக்குள் விருப்பத்தை வெளிப்படுத்தும்; நட்சத்திர குறி உள்ள குறியீடு பூஜியத்தையோ மீண்டும் மீண்டும் வருவதையே குறிக்கும். இந்த விதிகளின் வலது பக்க சீரான மொழிகளை விவரிப்பதாக எடுத்துக்கொள்ளலாம்.

{பெ.தொ, வே.தொ, வி, து.வி} ·

{(அ.அ), (அளவடை), பெ.அ * பெ} ·

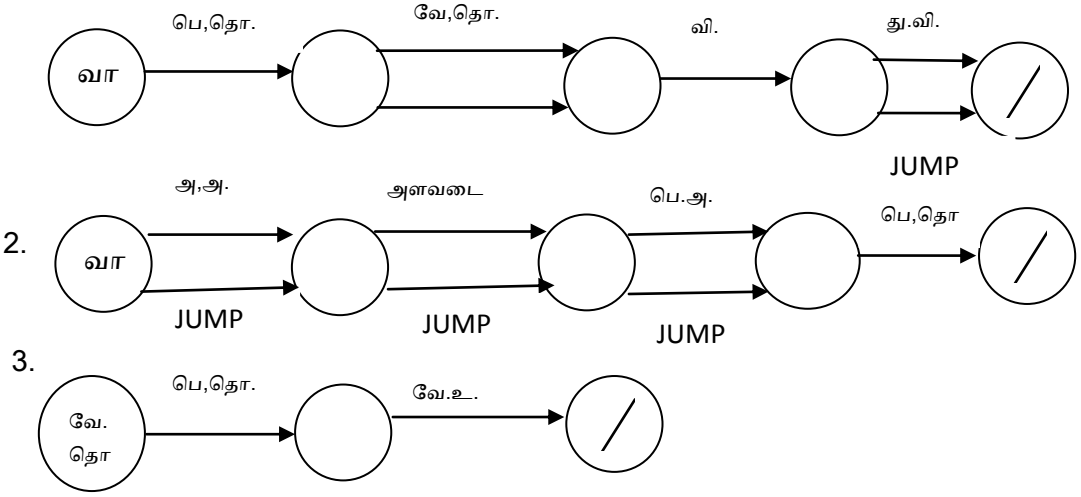
{பெ.தொ- வே.உ}

விதி 1-இன் இரு பகுதிகளுடன் தொடர்புடைய தானியங்கியின் நிலை வரைப்படங்கள் பின்வருமாறு அமையும்.



அ-வையும் ஆ-வையும் இணைத்து பின்வரும் 1. என்ற வரைபடம் கிடைக்கும். ஆக மேற்கண்ட நான்கு விதிகளுக்கு இணையாக பின்வரும் மூன்று நிலைமாற்ற வரைபடங்கள் கிடைக்கும்.

1.



Jumb என்பது எடிஎன் வடிவவாதத்தில் (ATN formalism) உள்ள எந்த உள்ளீட்டையும் ஏற்காத கட்டற்ற நிலைமாற்றத்தைக் (Unconditional transition) குறிப்பதற்கான இயல்பான மரபாகும். எளிய FSA-க்கு வகையான மறுதரவு செயல்பாட்டைக் கையாள இயலாது; ஏனென்றால் அது எங்கே இருந்து வந்தது எவ்வாறு உட்பட்டை கணித்தலுக்குப் பின் எங்குபோவது என்பதை நினைவு கூர்வதற்கு வழியில்லை. FSA-ஐ கீழ்தள்ளு அடுக்கை (Push down Stack) சேர்த்து கீழ் தள்ளு தானியங்கியாக (Pushdown Automata (PDA)) மாற்றினால் இது சாத்தியமானதாகும்.

அப்பொழுது இத்திட்டத்தில் கணித்தலின் சேமித்தல் மற்றும் மீட்டல் துணைச் செயன்மைகளைக் கையாளுவதற்குத் தேவையான மூலதனம் அமையும். கட்டுப்பாடு தற்காலிகமாக நிறுத்தப்பட்டு மீண்டும் தொடர்வதால் இது இறுதி உள்ளே–முதல்-வெளியே (last-in-first-out) என்ற அடிப்படையில் நிகழும். குறியீடுகளை (symbols) புலக்குறிப்புகளாகச் (Labels) செயல் அடிப்படையில் விளக்குதல் நிலைபெறு எடிஎன் (standard ATN) கருத்துச்சாயலில் பிரதிபலிக்கின்றது. உட்படுத்தும் செயற்பாங்கிற்கு கட்டளை தரப்பட்டால் அப்போதைய பொறிநிலை அடுக்கில் தள்ளப்படும். இவ்வாறு கட்டளைத் திரும்பும்போது கணித்தல் அதே இடத்திலிருந்து திரும்பத் தொடங்கப் பெறலாம். எடிஎன்வடிவவாதம் (ATN formalism) ஒரு நிலைக்கு கட்டுப்பாட்டின் பெயர்ச்சி எடுத்துக்காட்டாக PUSHNP (பெயர்த்தொடரைத் தள்ளு) என்ற அடையாளப்படுத்தப்பட்ட நிலைமாற்றத்தால் எழுப்பப்படும். இதற்குப் பொருத்தமான இறுதி நிலை மரபு POP என்று அடையாளப்படுத்தப்பட்ட பொய்யான வில்லை (arc) செயல்பாடு நிற்கும் போது மூலப் பொறிநிலை அடுக்கிலிருந்து திரும்பவும் தொடக்கநிலை நிறுத்தப்படும். நாம் இன்னும் பிற நிலைமாற்றங்களின் (து.வி, வி, அ.அ போன்றவற்றின்) மதிப்புநிலையை நிர்ணயிக்க வேண்டும். நமது விருப்பம் இலக்கண வகைப்பாட்டின் பெயர்களைப் பகுத்தாய்தலாக இருந்தால் ஒழிய அவை உண்மையான உள்ளீட்டுக் குறிகளாக இருப்பது அரிது.

நாம் நிலைமாற்ற வலைப்பின்னல் (Transition Network) இலக்கணங்கள் எவ்வாறு அகராதிகளைக் கையாளுகின்றன என்பதை குறிப்பிட வேண்டும். ஒரு சாத்தியம் அகராதியை விரித்தொழுது விதிகளின் (rewriting rules) தொகுதியாக வெளிப்படுத்த இயலும்.

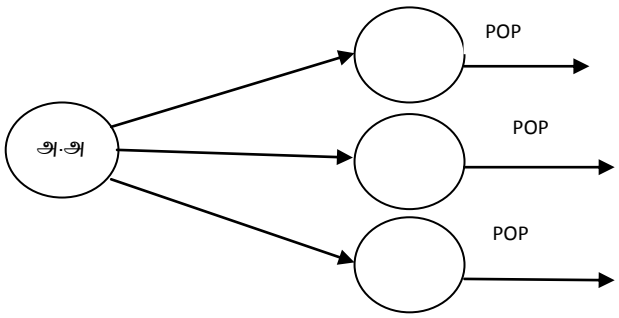
வி. 4. அ அ.அ→ ‘ஒரு’

வி. 4. ஆ அ.அ → ‘அந்த’

வி. 4. இ அ.அ → ‘கொஞ்சம்’

நாம் சில நுணுக்கங்களைப் பயன்படுத்தி விரித்தெழுது விதிகளை வடிவில் பின்வருமாறு மாற்றிக் கூறலாம்

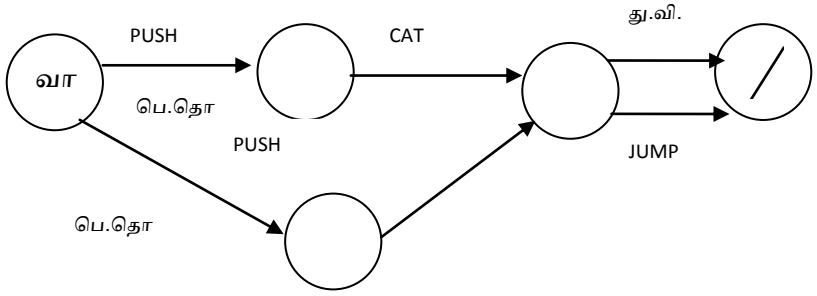
.



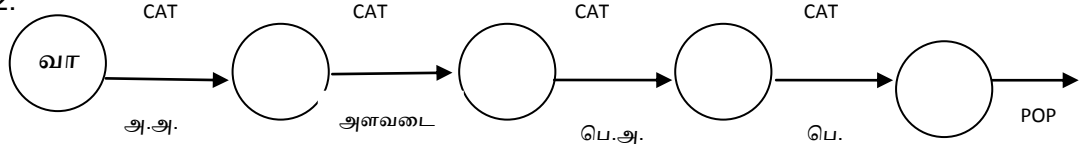
மாறாக ஏடிஎனில் எப்போதும் நடப்பது போல் நாம் நிலை மாற்றத்தைத் தூண்டும் ஒரு குழுமச்

செயல்பாடுகளை விவரிக்கலாம். அது முன் விவரிகாப்பட்ட அகராதியைப் பார்க்கும். அகராதியின் வடிவ அமைப்பு நிலைமாற்ற இலக்கணம் உள்ளடக்கம் கூடுதல் மொழிச் செயல்பாடாக மாறுகிறது. பரவலாகப் பயன்படுத்தப்படும் அகராதியைப் பார்க்கும் செயல்பாடுகள் (பொதுவாக CAT என்று அழைக்கப்படும்) முன் விளக்கம் பெற்ற வகைப்பாட்டை (Predifined category) அப்போதைய உள்ளீட்டுச் சொல்லுக்கான (Current input word) அகராதி சொல்லுடன் (பொருத்தி) பரிசோதிக்கின்றது. நாம் வலைப்பின்னல் இலக்கணத்தை (Network grammar) (A1-A3) சொல்வகைப்பாட்டுச் சோதனைகளுக்கிடையில் (CAT) மாற்றங்களை வேறுபடுத்துகின்ற மற்றும் (Push, தள்ளு) என்ற துணைச் செயல்பாடுகளை அழைக்கின்ற மற்றும் பொய் POP விற்களால் அடையாளப்படுத்தப்பட்ட இறுதிநிலைகளுடன் கூடிய ஆர்டிஎன்-க்கு நிலைபெறு கருத்துச்சாயலுடன் கூடுதல் நெருக்கமாகப் பொருத்தமுறுகின்ற ஒரு வடிவில் உருப்படுத்தம் செய்யும் நிலையில் நாம் இருக்கிறோம்.

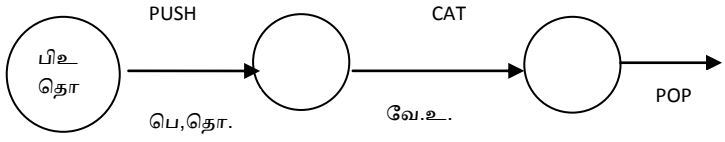
A1



A2:



A3:

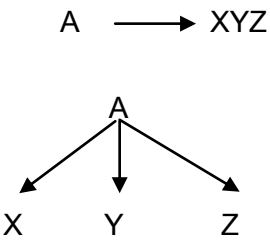


மேற்சொன்னது சூழல் வரையறை இல்லாத விரித்தொழுது விதிகளுக்குச் சமமான அருவநிலையாகும். அதாவது சூழல் வரையறை இல்லாத இலக்கணத்தை அதற்குச் சமமான ஆர்டிஎன்-ஆக மாறலாம். இருப்பினும் விரித்தொழுது விதி இல்லாத முக்கியமான ஆர்டிஎன் பண்புக்கூறு 'செயற்பாங்கு' (process) என்ற கருத்து சாயலாகும். நாம் ஒரு மொழியின் தெரிந்து

கொள்ளும் செயல்முறையை (recognition Process) உருவாக்க ஒரு இலக்கணத்தைப் பயன்படுத்த விரும்பினால் விரித்தொழுது விதி என்ன செய்ய வேண்டும் என்று கூறுகின்றது. ஆர்டிஎனில் அடங்கும் தானியங்கியும் எவ்வாறு செய்ய வேண்டும் என்று கூறுகின்றது. எ.கா. விரித்தெழுது விதி அடிப்படையிலான சூழல் வரையறையில்லாத 'தெரிந்துகொள்ளலுக்கு' உருவாக்கு ஒழுங்குமுறையும் (Production System) அதே மொழிக்கு 'தெரிந்து கொள்வானாக' நடைமுறைப்படுத்தப்பட்ட ஆர்டிஎனையும் ஒப்பிட்டுப்பார்க்கவும். ஆர்டிஎன் வடிவவாதம் (RTN Formalism) கணினி உள்ளீடுகளின் சில வரிசையான உருபடுத்தம் செய்யும் போது மரபு செயல்முறை நிரலாக்க மொழிகளின் (Conventional Procedured Programming Languages) பண்புகளை வெளிப்படுத்துகின்றது: கட்டுண்டவைகள் மற்றும் கட்டுப்படாதவைகள் இவற்றின் 'go to' நிலை மாற்றங்களும் மொழிபு அடையாள குறிகள் (நிலையின்பெயர்கள்), செயற்பாங்கு கட்டளைகள் (Procedurce calls) (Push: POP). இந்தக் கருத்து சாயலில் முக்கியமான வெளிப்படையாக இல்லாத செயற்பாங்கு 'விளங்கிக்கொள்வான்' (Interpreter) அதிலிருந்து ஒன்றுக்கும் மேற்பட்ட நிலைமாற்றம் வரும் நிலையை அடைகையில் அதன் நடத்தை ஆகும்.

5.2.1.2.3.2. நிலைபெறு பெரிதாக்கப்பட்ட நிலைமாற்ற வலைப்பின்னல்கள் (Standard ATNS)

ஆர்டிஎன் (RTN) பகுத்தாய்வுக்கு மிகப் பயனுள்ள பொறி அல்ல. இது சூழல் வரையறையில்லாத மொழிகளை மட்டும் தான் தெரிந்துகொள்ளும். அமைப்பாக்கம் செய்யப்பட்ட வெளியீட்டைத் தருவதற்கான வழிமுறை இல்லை. உறுப்பமைப்பு இலக்கணங்கள் என்று அறியப்படுகின்ற விரித்தொழுது விதி ஒழுங்குமுறைகளின் ஓரளவுக்குப் பெரிய துணைவகுப்புக்கு விரித்தெழுது அடையாளக் குறிகளைப் படிநிலை ஆதிக்க இயக்கியாக (Hierachical dominece operator) மீட்டுருவாக்கம் செய்து ஒரு புரிந்து கொள்ளுதல் செயல்பாட்டிலிருந்து வெளியீடாக ஒரு பகுத்துக்குறித்தல் கிளை அமைப்பை உருவாக்க இயலும்.



ஆர்டிஎன்னை (RTN) சூழல் கட்டுண்ட பகுத்துக்குறித்தலாகப் பயன்படுத்த கடுமையாகத் திட்டமிடல் வேண்டும். வடிவவாதம் ஒரே நேரத்தில் ஒன்றுக்கும் மேற்பட்ட குறியீடுகளைப் (Symbols) பரிசோதிக்க வழி தராத காரணத்தால் கட்டுண்ட தொடரமைப்பு விதிகளின் (Context Sensitive Phrase Structure rules) பின்வரும் நிலைபேறு வடிவங்களில் ஒன்றை நிலைமாற்ற

வலை அமைப்பு (Transition Network) உருப்படுத்தமாகப் பதிலீடு செய்ய நேரடியான வழி இல்லை.

XAY ⟶ X BC Y

அல்லது

A ⟶ BC/X–Y

முற்றுநிலை தெரிந்து கொள்வானின் (Finite State Recognizer (FSR)) கரடுமுரடான கட்டமைப்பிலிருந்தான (crude arctitecture) மாதிரி வடிவில் பரம்பரையாகப் பெற்ற விளக்கங்களின் பரம்பரை மேல்-கீழ் (top-down) இடமிருந்து வலம் (left-right தொடர்ச்சி நுணுக்கங்களின் இவ்வகையிலான விதிகளை நினைத்துப்பார்ப்பதைக் கூட நிலைமாற்ற வலை அமைப்புச் சட்டகத்தில் விகாரமாகச் செய்கிறது. சூழல் கட்டுண்டமையின் மேல்-கீழ் இடமிருந்து வலம் நடைமுறை எது பின்னால் வருகிறது என்று அறியும் சந்தர்ப்பம் கிடைப்பது வரை நிச்சயங்களைப் பின்தள்ளிவைக்கச் சாத்தியப்பட வேண்டும். நிலைமாற்ற வலை அமைப்பின் இயந்திரத்தன்மையான பரிமாணம் அடிப்படையில் பின்னால் வரும் சில நிலை மாற்றங்களுக்குக் (transition) கட்டுப்பாடுகளைக் கூடுதலாகத் தர முந்திய நிலைமாற்றங்களின் விவரணைகளைச் சேமித்துவைக்கவும் மீட்கவும் கூடுதலான நினைவகம் தேவை. எடிஎன்னில் (ATN) வலை அமைப்பின் ஒவ்வொரு நிலையிலும் ஒரு குழுமப்பதிவுகளை அவற்றை பயன்படுத்தும் வண்ணம் சில முன்விளக்கப்பட்ட செயல்பாடுகளுடன் சேர்க்கின்றோம். நிலைபேறு எடிஎன்னில் (ATN) பதிவுகளில் தரப்பட்டுள்ள மிக முக்கியமான செயன்மைகள் (operations) SETR என்ற ஒப்படைக்கும் செயல்பாடும் ஒரு பதிவுக்கு தற்போதைய மதிப்பைத் தரும் GETR என்ற செயல்பாடும் ஆகும். பதிவுகளையும் (registers) அது தொடர்பான செயன்மைகளையும் சேர்த்து ஆர்டிஎன்னை சூழல் கட்டுண்ட மொழிகளைத் தெரிந்துகொள்ளச் செய்யலாம். இருப்பினும் இலக்கணம் வழியாகப் பயணத்தின் தடத்திலிருந்து தானாகவே பகுத்தாய்ந்த கிளை அமைப்பை (parse tree) உருவாக்கும் முக்கியமான பண்பை இழந்துவிடலாம். இருப்பினும் ATN literals மற்றும் பதிவுப்பொருளடக்கத்திலிருந்து எந்த பட்டியல் அமைப்பை உருவாக்குவதை அனுமதிக்கும் அமைப்பு-உருவாக்குச் செயல்பாடான (structure bullding fundion) BUILD 2-ஐத் தருகிறது. மேலும் இந்த இடுகுறித்தன்மையான பட்டியல் கிளை அமைப்புகளாகப் புரிந்துகொள்ளப்படுகின்றன. கிளை அமைப்பின் எந்த வகையையும் காட்டவும் சேகரிக்கவும் BUILD 2-இன் பயன்பாட்டால் சாத்தியமாகும். இதை வெளியீடு செய்யவும் வசதி வேண்டும். எடிஎன் கட்டளைச் செயல்பாட்டிற்கு (calling process) ஒரு மதிப்பைத் திருப்பித் தர POP-ஐ அனுமதித்து வெளியீட்டைத் தருகிறது. பதிவுகள் வெளியீடு மற்றும் பதிவுகளின் மேல் சோதனைகளால் நிலைமாற்றங்களைக் கட்டுப்படுத்தும் ஆற்றல் இவற்றில் ஆர்டிஎன் (RTN) எடிஎன் (ATN) ஆக மாறுகிறது. சொற்களின் தொடர்ச்சியை கிளை அமைப்பாக மாற்றுகிறது.

செயல்முறைக் கட்டளைகள்

1. (.CAT.): வகைப்பாட்டுப் பரிசோதனையைச் செயல்படுத்தும் காரியக் கிரமம்.

2. (.PUSH.): தள்ளுப் பரிசோதனையைச் செயல்படுத்தும் காரியக் கிரமம்.

3. (.TST.) : நிலைமாற்ற அட்டவணைப் பரிசோதனையைச் செயல்படுத்தும் காரியக் கிரமம்

4. (.POP.): இறுதியைத் தீர்மானித்து Bருடுனு ஐக் கூப்பிடும் காரியக் கிரமம்.

5. (.SETR.): பதிவின் மதிப்பை நிறுவும் காரியக் கிரமம்.

6. (.SENDR.): கீழ்நிலைக்கு பதிவு மதிப்பை அனுப்பும் காரியக் கிரமம்.

7. (.CIFTER.): மேல்நிலைக்கு பதிவு மதிப்பைத் திருப்பும் காரியக் கிரமம்.

8. (.TO.): உள்ளீட்டுக் காட்டியை முன்னேற்றி நடைமுறைபடுத்த வேண்டிய அடுத்த கணுவை நிர்ணயிக்கும் காரியக்கிரமம்.

9. (.JUMP.): நடைமுறைப்படுத்த வேண்டிய அடுத்த கணுவை நிர்ணயிக்கும் காரியக் கிரமம்.

10. (.GETR.): பதிவின் மதிப்பை பெறும் காரியக் கிரமம்.

11. (ooo): நட்சத்திரப் பதிவின் மதிப்பை பெறும் காரியக்கிரமம்.

12. (.GETF.): அகராதியிலிருந்து பண்புக்கூறை மீட்டும் காரியக் கிரமம்.

13. (.APPEND.): ஒரு தரவு சாதனத்தைப் பதிவில் சேர்க்கும் காரியக் கிரமம்

14. (.BUILD.): புதிய அமைப்பை உருவாக்க வேண்டி பல பதிவுகளில் இருக்கும் மதிப்புகளை இணைக்கும் காரியக் கிரமம்.

5.2.2. பாணினி இலக்கண வடிவவாதம்

பாணினி இலக்கண வடிவவாதம் இந்திய மொழிகளைப் பகுப்பாய்வு செய்வதற்காக அறிமுகப் படுத்தப்பட்ட வடிவவாதம் ஆகும் (Akshar Bharathi மற்றும் பிறர் 1995). இவ்வடிவவாதம் தொடரமைப்பு இலக்கணம் (phrase structure grammar), சூழல் கட்டில்லா இலக்கணம் (Contex free grammar), சூழல் கட்டுண்ட இலக்கணம் (context sensitive grammar), ஆக்கமுறை இலக்கணம் (generative grammar), டாக் இலக்கணம் (Tag Grammar), சொல்சார் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG)) போன்ற இலக்கண வடிவவாதங்கள் ஆங்கிலம் போன்ற SOV அமைப்புள்ள மொழிகளுக்குத்தான் பொருந்தும் என்ற வாத்தத்தை முன்வைத்து இந்திய மொழிகளுக்காக முன்மொழியப்பட்டது.

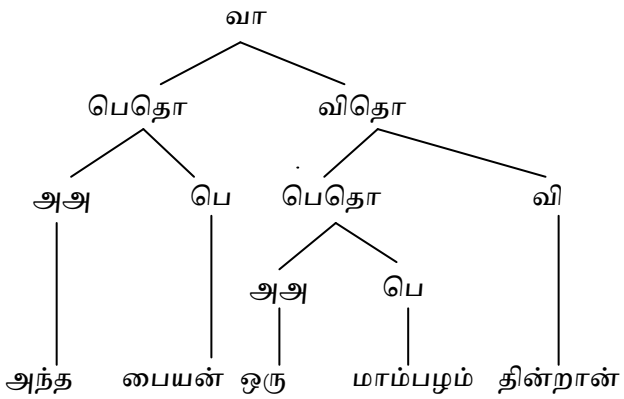
இலக்கண வடிவவாதம்

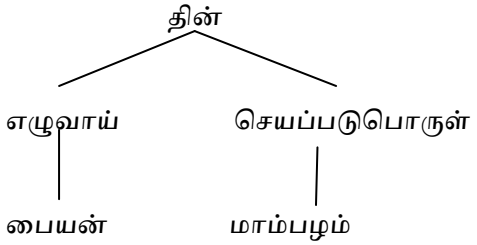
இந்திய மொழிகளின் பகுப்பாய்வுக்கும் பகுத்துக் குறித்தலுக்கும் பாணினியின் வடிவவாதம் மிகப் பரவலாகப் பயன்படுத்தப்படுகின்றது. இவ்வடிவ வாதத்தார் இந்திய மொழிகளுக்குப் பிற வடிவவாதங்களை விட பாணினியின் வடிவவாதம் பொருத்தமானது என

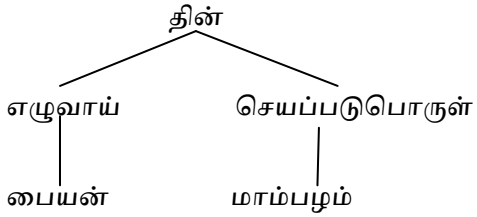
வாதிடுகின்றனர். கூற்றுக்கும் கேட்பவர் எடுத்துக் கொள்கிற பொருண்மைக்கும் இடையே உள்ள உறவை நிலைநிறுத்தும் விதிகளின் ஒழுங்குமுறை (இயற்கை மொழியின்) இலக்கணமாகும். எடுத்துக்காட்டாக பின்வரும் வாக்கியத்தை எடுத்துக்கொள்வோம்.

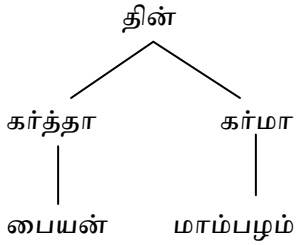
அந்தப் பையன் ஒரு மாம்பழம் தின்றான்

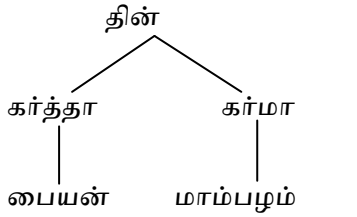
வழி 1

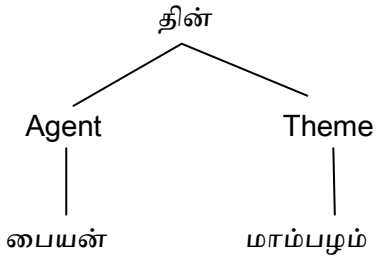


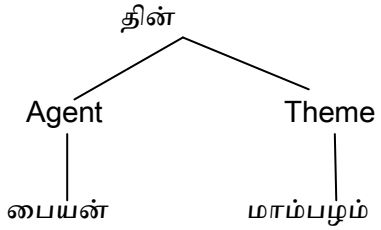












இலக்கண வடிவவாதம் 2

விரும்புகிற பொருளைத் தருகிற சொற்குழுக்களுக்கு இடையே உறவைப் பிரித்தெடுக்கும் நுணுக்கத்தை தருவது இந்த இலக்கண வடிவவாதத்தின் வேலையாகும். பாணினி இலக்கண வடிவவாதம் பின்வரும் முக்கியமான கேள்வியை எழுப்பபுகின்றது: கூற்றிலிருந்து பொருண்மையைக் கேட்பவர் எவ்வாறு பெறுகிறார்? அல்லது எவ்வாறு மொழி பொருண்மையைக் குறியாக்கம் செய்கின்றது?

- எந்த வடிவில்?

- எங்கே?

– எவ்வளவு?

பாணினி இலக்கணத்தின் பொருண்மை மாதிரி பாணினி இலக்கணத்தின் பொருண்மை மாதிரி செயலைச் செய்கையாகவும் விளைவாகவும் பகுக்கின்றது. செயல் செய்கை விளைவு பாணினியின் பொருண்மை மாதிரியில் முக்கியமான கருத்துச்சாயல் காரகம் ஆகும். காரகம் செயலுக்கும் அதில் பங்கு பெறுவோர்க்கும் இடையே உள்ள உறவை வெளிப்படுத்தும் இம்மாதிரியில் செய்கையானது பல துணைச் செய்கைகளாகப் பிரிக்கப்படும்.

செய்கை:

துணைச் செய்கை 1

துணைச் செய்கை 2

துணைச் செய்கை 3

துணைச் செய்கை 4

பேசுபவரின் விருப்பம் இம்மாதிரியில் வெளிப்படும். இம்மாதிரி பின்வரும் பண்புக் கூறுகளால் இந்திய மொழிகளுக்குப் பொருத்தமாக அமைகின்றது.

1. சுத்தமான சொல் வருகை முறை

2. வேற்றுமை உருபுகள் காரணமாக இருத்தல்

பெயருடனான வினையின் உறவு வேற்றுமை உருபுகளாகும். வாக்கியத்திலிருந்து எவ்வாறு காரக உறவுகளைப் பிரித்தெடுப்பது? என்ற கேள்விகளுக்கு விடை இரண்டு சிக்கல்களை நேரிடும். முதல் சிக்கல்

வேற்றுமை உருபு-காரகம் பல-ஒன்று பொருத்தம் காட்டும்

எடுத்துக்காட்டாகப் பின்வரும் வாக்கியங்களைக் கருத்தில் கொள்ளவும்.

1. இராமன் மாங்காய் தின்கிறான்

2. இராமன் மாங்காயைத் தின்கிறான்

3. இராமனுக்கு மாங்காய் தின்ன வேண்டும்

4. இராமனால் மாங்காய் தின்னப்பட்டது

5. இராமனால்/ இராமனுக்கு மாங்காய் தின்னமுடியவில்லை

மேற்சொன்ன எல்லா வாக்கியங்களிலும் இராமன் - கர்த்தா (எழுவாய்) ஆனால் வேற்றுமைகள் வேறுபட்டவை. தீர்மானம் வேற்றுமை உருபுகள் செயல்பாடு (வினையின் காலம், ஆற்றல்) நோக்கு சார்ந்தது. குறிப்பாக வேற்றுமை உருபு-காரகப் பொருத்தம் வினையிலிருந்து சுதந்திரமானது மற்றும் வினையின் காலம், ஆற்றல், நோக்கு இவற்றைச் சார்ந்தது. உருப்படுத்தம் முன்னர் தரப்பட்ட எடுத்துக்காட்டு அடிப்படையில் தின் என்ற வினைக்கான காரக அட்டவணை பின்வருமாறு அமையும்.

எடுத்துக்காட்டு

அ) தின் என்ற வினைக்கான காரக அட்டவணை

காரகம்	தேவை	வேற்றுமை உருபு
கர்த்தா	கட்டாயம்	0
கர்மா	கட்டாயம்	0/ஐ
காரணம்	விருப்பம்	ஆல்

(ஆ) மாற்று விதிகள் வினையின் காலம், ஆற்றல், நோக்கு அதன் அடிப்படையில் மாற்றங்கள் ஏற்படும் மாற்றங்கள்

அ - வேண்டும் கர்த்தா : 0 > உக்கு

அ - பட்டது கர்த்தா : 0 > ஆல்

அ - முடியவில்லை/ முடியும் கர்த்தா : 0 > உக்கு/ ஆல்

இரண்டாவது சிக்கல் வேற்றுமை உருபுக்கும் காரகத்திற்கும் இடையில் ஒன்று பல பொருத்தம் வரும்.

எ.கா.

ராமன் சாவியால் பூட்டைத் திறந்தான்

சாவி பூட்டைத் திறந்தது

பூட்டு திறந்தது

வேற்றுமை உருபு + 0 மேற்சொன்ன எடுத்துக்காட்டில் எழுவாயின் காரகம், கர்த்தா, காரணம், கர்மா என்ற மூன்று நிலைகளில் வருகின்றது. எந்தச் செயலின் காரகம் என்பது இச்சிக்கலுக்குத் தீர்வைத் தரும். தீர்மானம் அகப் பொருண்மை மாதிரி பின்வருமாறு பகுத்தாயப்பட வேண்டும். பூட்டுத் திறத்தல் (முக்கிய செயல்)

1. சாவியைச் செருகித் திருப்புதல்

2. சாவியை நெம்புகோலில் அழுத்தி அதை நகரச் செய்வது

3. தாழ் பின் வாங்குதல் மற்றும் பூட்டுத் திறத்தல்

ராமன் : தொடக்கச் செயல்பாடு ஒன்றின் காரகம்

சாவி : தொடக்கச் செயல்பாடு இரண்டின் காரகம்

பூட்டு : தொடக்கச் செயல்பாடு மூன்றின் காரகம்

காரகப் பங்களிப்புகளைத் தருவதில் வாக்கிய கட்டுப்பாடுகள் பின்வருமாறு அமையும்:

1. ஒரு வினைக் குழுவின் ஒவ்வொரு கட்டாயக் காரகமும் சரியாக ஒரு மூலச்சொல் குழுவுக்குத் தரப்பட வேண்டும்.

2. ஒரு வினைக் குழுவின் ஒவ்வொரு விருப்புக் காரகமும் சரியாக குறைந்தது ஒரு மூலச் சொல் குழுவுக்குத் தரப்பட வேண்டும்.

3. ஒவ்வொரு மூலச் சொல் குழுவும் சில வினைக்குழுவில் சரியாக ஒரு காரகத்திற்குத் தரப்பட வேண்டும்.

பில்மோர் (Fillmore, 1968) காரகக் கோட்பாட்டைப் பின்வரும் நான்கு நிலைகளாகக் குறிப்பிடுகிறார். அவையாவன:

பொருண்மை நிலை (பேசுபவர் மனதில்)

காரக நிலை வேற்றுமை

உருபு நிலை (வேற்றுமை உருபு)

புறநிலை (வாக்கியம்)

எடுத்துக்காட்டு, பொருண்மை நிலை

தின் : செயல் ராமன் R1, மாங்காய்: R2, கை : R3

காரக நிலை தின் - இறந்தகாலம், ராமன் : K1, மாங்காய் K2, கை : K3

வேற்றுமை நிலை (வேற்றுமை உருபு)

தின் : த் ராமன் : மாங்காய் : கை: ஆல்

புற நிலை (வாக்கியம்)

5.2.2.1.பாணினி பகுத்தாய்வான் (Paninian Parser)

பாணினியின் பகுத்தாய்வான் அடிப்படையில் ஒரு மொழியின் சொற்குழுமங்களையும் இலக்கணத்தையும் பயன்படுத்தி முக்கியப் பகுத்தாய்வான் வாக்கியத்தின் பகுத்தாய்வைச் செய்யும் இதைப் பின்வருமாறு உருப்படுத்தம் செய்யலாம்.

மொழி L-இலிருந்து சொற்குழுக்கள்

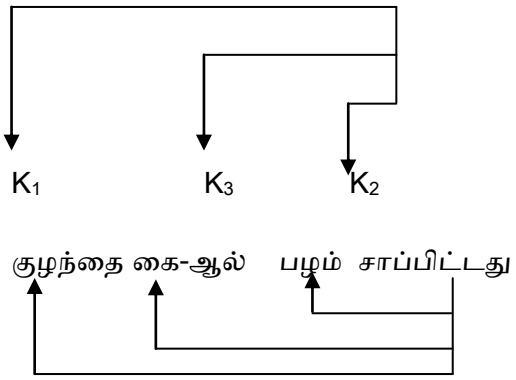
மொழி L-இன் இலக்கம்

முக்கியப் பகுத்தாய்வான் வாக்கியத்தின் பகுத்தாய்வுக்கு பின்வருவனவற்றை வேண்டும்.

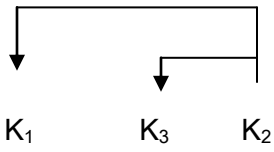
1.(அ) வினை வேர்களுக்கான காரக அட்டவணை

(ஆ) மாற்று விதிகளின் காரக அட்டவணை

2.காரகப் பங்களிப்பு தரப்படுவதன் மேலான வாக்கிய கட்டுப்பாடுகள் ஆகியவையாகும். காரக அட்டவணையைப் பயன்படுத்தி நாம் பின்வரும் கட்டுப்பாடு வரைபடம் பெறுகின்றோம்.



வாக்கிய கட்டுப்பாடுகளைப் பயன்படுத்தி நாம் இரண்டு தீர்மான வரைபடங்களைப் பெறுகின்றோம்.



குழந்தை கை-ஆல் பழம் சாப்பிட்டது

K_1	கர்த்தா
K_2	கர்மா
K_3	காரணம்

5.2.2.1.1. முக்கியப் பகுத்தாய்வான் (Core Parser)

முக்கியப் பகுத்தாய்வான் பின்வரும் செயல்பாடுகளைச் செய்யும்:

1) சொற்களுக்கு இடையில் காரக உறவை அடையாளம் காணும் (எழுவாய்க்கும், வினைக்கும் இடையே உள்ள உறவு)

2) சொற்களின் பொருண்மைகளை அடையாளம் காணும்

1. காரக- வேற்றுமை உருபு பொருத்தம்
2. காரகத்தின் விருப்பம்
3. மாற்று விதிகள்

5.2.2.1.2. குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் (Local Word Grouping)

இது பாணினி பகுத்தாய்வானில் முக்கியமான கருத்துச் சாயலாகும். இதன் படி சொற்கள் பகுத்தாய்தலுக்காக ஒன்றாகக் குழுமப்படும். ஆனால், இச்செயல்பாடு தொடரமைப்பு இலக்கணங்களில் செய்யப்படும் 'வாக்கியங்களைத் தொடர்களாகப் பிரித்தல்' என்பதிலிருந்து பின்வரும் உண்மைகளின் அடிப்படையில் வேறுபடும்.

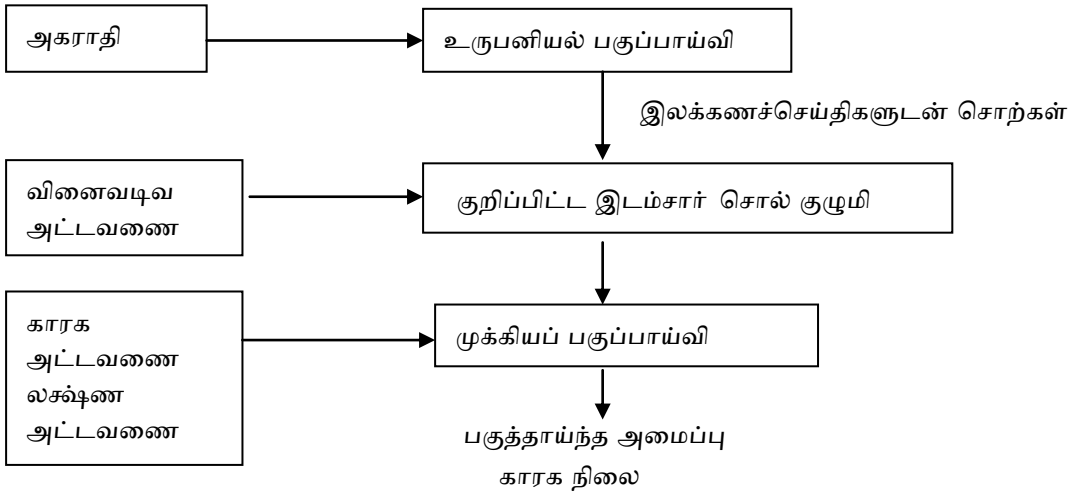
- இந்திய மொழிகளில் வினைத்தொடர் இயற்கையாக அமையத் தேவையில்லை.
- கணித்தல் அடிப்படையில் பெயர்த் தொடர்களையும் வினைத் தொடர்களையும் தெரிந்து கொள்ளுதல் அவ்வளவு எளிதானதோ வலுவானதோ அல்ல.
- எனவே பெயர்க்குழுக்களும் வினைக் குழுக்களும் புறச் செய்திகளைக் கொண்டு செயல்படுகின்றது.
- முக்கியமானது வினைக் குழுக்களின் பொருண்மைச் சாயல்களின் தெரிதலாகும்.

பின்வருவன முக்கியமான தகவல்களாகும்.

1. வினை வேர்களின் சாத்தியமான தொடர்ச்சி
2. பால்-எண்-இட உடன்பாடு

இதுவரைச் சொன்ன செய்திகளின் அடிப்படையில் பாணினி பகுத்தாய்வானைப் பின்வருமாறு உருப்படுத்தம் செய்யலாம்.

வாக்கியம்



5.2.2.2.பகுத்தாய்வின் முக்கியத்துவம்

பகுத்தாய்தல் காரகங்களின் முக்கியத்துவம் பின்வருமாறு அமையும்:

- கர்த்தாவுக்கு கீழ்நோக்கு வரிசையில் பின்வருமாறு முக்கியத்துவம் அமையும்: மனிதன், மனிதன் அல்லா உயிர்கள், உயிரில்லாதவை
- மூலக் குழுமம் அது விரும்புகிற குழுமத்துடன் சேர்ந்து வரும்
- கர்த்தா கர்மாவுக்கு முன்வரும்.

பொருள் மயக்கத்தைப் போக்க லக்ஷ்ண அட்டவணை வாக்கியத்தில் உள்ள வினையின் பொருண்மையைக் காண வினைகளின் காரகத்தைப் பயன்படுத்துகின்றது.

எ.கா.

அவன் அவளிடம் மன்னிப்புக் கோரினான்

அவன் கிணற்றிலிருந்த தண்ணீர் மொண்டான்

அவன் அவளிடம் பணம் கேட்டான்

அவன் அவள் சொல்வதைக் கேட்டான்

வினை லக்ஷ்ண அட்டவணை அகராதிகளைப் பயன்படுத்தி மொழியியலாளர்களால் உருவாக்கப்படும் பெயர் லக்ஷ்ண அட்டவணை ஒரு வாக்கியத்தில் வரும் பெயர்களின் பொருண்மை மயக்கங்களைப் போக்க உதவும்.

எடுத்துக்காட்டு

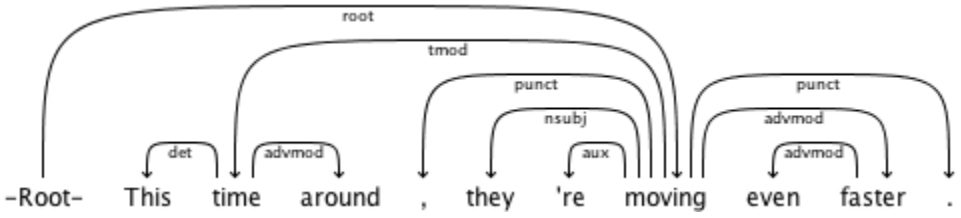
அவன் வயலை உழுகிறான்

அவன் வயலை அறுக்கிறான்

5.2.3. சார்புப் பகுப்பாய்வு (dependency parsing)

ஒரு சார்பு பாகுபடுத்தி (dependency parser) ஒரு வாக்கியத்தின் இலக்கணக் கட்டமைப்பைப் பகுப்பாய்வு செய்து, "தலை" சொற்களுக்கும் அந்தத் தலைகளை அடைசெய்யும்

சொற்களுக்கும் இடையிலான உறவுகளை நிறுவுகிறது. கீழே உள்ள படம் ஒரு குறுகிய வாக்கியத்தின் சார்பு பாகுபாட்டைக் காட்டுகிறது. moving என்ற வார்த்தையில் இருந்து faster என்ற வார்த்தைக்கு நகரும் அம்புக்குறி moving என்பதை faster அடை செய்கின்றது என்று காட்டுகின்றது. மேலும் அம்புக்கு ஒதுக்கப்பட்ட advmod என்ற புலக்குறிப்பு சார்புநிலையின் சரியான தன்மையை விவரிக்கிறது.



moving என்பது வேராகக் (root) கருதப்பட்டு root எனப் புலக்குறிப்பு செய்யப்பட்டுள்ளது; this என்பது time என்பதன் அடைகொளி அடையாகக் கருதப்பட்டு det என புலக்குறிப்பு செய்யப்பட்டுள்ளது; around என்பது time என்பதன் வினையடையாகக் கருதப்பட்டு advmod எனப் புலக்குறிப்பு செய்யப்பட்டுள்ளது; they என்பது moving என்பதன் எழுவாயாகக் கருத்தப்பட்டு nsub எனப் புலக்குறிப்பு செய்யப்பட்டுள்ளது. faster என்பது moving என்பதன் வினையடையாகக் கருதப்பட்டு அதனுடன் அம்புக்குறியால் இணைக்கப்பட்டு advmod எனப் புலக்குறிப்பு செய்யப்பட்டுள்ளது; even என்பது faster என்பதன் வினையடையாகக் கருதப்பட்டு அதனுடன் அம்புக்குறியால் இணைக்கப்பட்டு advmod எனப் புலக்குறிப்பு செய்யப்பட்டுள்ளது; ‘re (are) என்பது moving என்பதன் துணைவினையாகக் கருத்தப்பட்டு அதனுடன் அம்புக்குறியால் இணைக்கப்பட்டு aux எனப் புலக்குறிப்பு செய்யப்பட்டுள்ளது. time என்பது moving என்பதன் கால அடையாகக் கருத்தப்பட்டு அதனுடன் அம்புகுறியால் இணைக்கப்பட்டு tmod என அடையாளப்படுத்தப்பட்டுள்ளது. நிறுத்தற்குறி (.) moving என்பதுடன் அம்புக்குறியால் இணைக்கப்பட்டு punct எனப் புலக்குறிப்பு செய்யப்பட்டுள்ளது.

இயற்கையான மொழி வாக்கியங்களின் வகைப்படுத்தப்பட்ட சார்புப் பகுப்பாய்வுகளை உருவாக்கும் மிக விரைவான நிலைமாற்றம்-அடிப்படையிலான பாகுபடுத்தி ஸ்டன்போர்டு இயற்கை மொழி ஆய்வுக் குழுவால் (Standford Natural Language Processing Group) உருவாக்கப்பட்டுள்ளது. ஒரு நரம்பியல் வலைப்பின்னலால் (neural network) இயக்கப்படும் இந்த பகுப்பாய்வி உட்பொதித்தல் (word embedding) உள்ளீடுகளை (inputs) ஏற்றுக்கொள்ளும்.

இந்த பாகுபடுத்தி (உலகப்பொது சார்புநிலைகள், ஸ்டான்போர்ட் சார்புநிலைகள் மற்றும் கோஎன்எல்எல் சார்புகளுடன்) ஆங்கிலத்தையும் சீன மொழியையும் (கோஎன்எல்எல்

சார்புகளுடன்) ஆதரிக்கிறது. மென்பொருளின் எதிர்காலப் பதிப்புகள் பிற மொழிகளை ஆதரிக்கும்.

5.3. ஆழமில்லாப் பகுப்பாய்வு

நாம் ஆழமில்லாப் பகுப்பாய்வு (shallow parsing) என்ற கலைச்சொல்லை மரபுப் பகுப்பாய்விலிருந்து கிடைக்கும் வெளிப்பீட்டை விடக் குறைவான நிறைவுடைய ஆய்வைக் குறிக்கும் பொதுவான சொல்லாகப் பயன்படுத்துகிறோம். ஆய்விலிருந்து கிடைக்கும் வெளிப்பீடு தொடரமைப்புக் கிளைப்படம் அல்ல. ஒரு பகுப்பாய்வி பெயர்த்தொடர்கள் போன்ற பகுக்கப்பட்ட உறுப்புகளை வாக்கியத்தில் அவற்றின் அக அமைப்பையும் செயல்பாட்டையும் குறிப்பிடாமல் அடையாளம் காணும். மற்றொரு வகையான ஆய்வு முதன்மை வினை மற்றும் அவற்றின் நேரடியான பங்கெடுப்பாளர் போன்ற சொற்களின் சில செயல்பாட்டுப் பங்களிப்பை அடையாளம் காணும். பகுப்பாய்வு ஒழுங்குமுறை உருபயனியல் ஆய்வின் மீதும் பொருண்மை மயக்க நீக்கலின் மீதும் இயல்பாகச் செயல்படும். இதன் அடிப்படை நோக்கம் தலைச்சொல் (Head word) உருபனியல் தகவல் மற்றும் சொல்வரிசை அமைப்பு இவற்றில் இருந்து சாத்தியமான தொடரியல் அமைப்பைப் (Syntantic structure) பெறுவதாகும். முன்மாதிரியாக பகுப்பாய்வுத் தொடர்களையும் தலை-அடை (head-medifier) உறவுகளையும் கண்டுபிடிப்பதை நோக்கமாகக் கொள்கிறது. பெரிய உரை தரவுத்தொகுதிகளின் (corpus) பயன்பாட்டில் ஆழமில்லாப் பகுப்பாய்வி பயன்படுத்தப்படுகிறது. பெரும்பாலும் பகுப்பாய்விற்கு எல்லாச் சிக்கல்களுக்கும் தீர்வு காண்பதற்கு போதிய திறன் இல்லாவிட்டால் பகுதியான ஆய்வுகள் அனுமதிக்கப்படும். சர்ச் (Church.1988) பொருத்தமான அடைப்புக் குறிகளைச் செருகி எளிய பெயர்த் தொடர்களை அடையாளம் காணும் புள்ளியல் (Stochastic) வழியமைப்பை வடிவமைத்துள்ளார். ஒப்னெய் (Obney, 1991) உளமொழியியல் சான்றுகள் அடிப்படையிலும் நடைமுறை பயன்பாடுகள் அடிப்படையிலும் ஆழமில்லாப் பகுப்பாய்வின் முக்கியத்துவத்தை முதலில் விவாதித்தவர் என்று மதிக்கத்தக்கவர் ஆவார். அவர் தமது அணுகுமுறையில் ஆழமில்லாப் பகுப்பாய்வைப் பெறுவதற்காக முற்றுநிலை மாறிகளைப் பயன்படுத்தினார். ஆழமில்லாப் பகுப்பாய்வி சொல்வகைப்பாடு அடையாளப்படுத்தல், தொடர் உறுப்பு வரையறுத்தல், தொடர் உறவுகளைக் கண்டுபிடித்தல் என்ற முன்மாதிரி தொகுப்புகளை உள்ளடக்கியது.

5.3.1 சொல்வகை அடையாளப்படுத்தல்

ஒரு சொல்லும் அதன் சூழலும் தரப்படுகையில் சொல்வகை அடையாளப்படுத்தி அச்சொல்லின் சரியான உருபனியல்-தொடரியல் வகுப்பை (பெயர், வினை கொண்ட பகுப்பை)

உறுதிசெய்யும் சொல்வகை அடையாளப்படுத்தி சொல்வகை அடையாளப்படுத்தல் (Parts of Speech tagging) இயற்கைமொழி ஆய்வில் மிக அறியப்பட்ட சிக்கலாகும். இதில் இயந்திர கற்றல் அணுகுமுறைகள் (machine language approaches) பொதுவாகப் பயன்படுத்தப்படுகிறது.

5.3.2. தொடர் பகுத்தல்

சொற்களும் அவற்றின் உருபனியல்-தொடரியல் வகுப்புகளும் தரப்படுகையில் எந்தச் சொற்கள் தொடர்களாகக் (பெயர்த்தொடர், வினைத்தொடர், நிரப்பியத்தொடர் போன்றவைகளாக) குழுமப்பட வேண்டும் என்று தொடர் பகுத்தல் (Chunking) செயல்பாட்டால் தீர்மானிக்கப்படுகிறது.

தொடர் பகுத்தலுடன் பின்வரும் கேள்வி தொடர்புள்ளது. தொடர் எதைக் கொண்டிருக்கும்? ஒரு மாதிரித் தொடர் ஒரு தனி நிலைப் பொருளடக்கச் சொல்லைச் சுற்றி ஒரு கூட்டம் செயல் பாட்டுச் சொற்களைக் கொண்டிருக்கும் (Abney, 1991). தொடர்கள் இயல்பாக ஒன்றுக்கொன்று இயைபுள்ள சொற்களின் குழுமமாக எடுத்துக்கொள்ளப்படுகின்றது.

அடுத்த கேள்வி 'நம்முடைய தேவைக்காக இந்த இயைபுள்ள சொற்களின் குழுவின் எல்லைகளை எவ்வாறு வரையறை விளக்கம் செய்யவியலும்?' என்பதாகும். எடுத்துக்காட்டாக, பின்வரும் இணைகளில் எது சரியான தொடர்? என்ற ஐயம்.

1. ((மேசையின் மேல்)) அல்லது
2. ((மேசையின்)) ((மேல்))

5.3.3. தொடர் உறவைக் கண்டுபிடித்தல்

ஒரு வாக்கியத்தில் உள்ள தொடர்கள் தரப்படுகையில் முக்கிய வினையுடன் அவற்றிற்குள்ள உறவுகள் (எழுவாய், பயனிலை, இடம் போன்றவை) தீர்மானிக்கப்பட வேண்டும் என்ற நிலையில் உறவைக் கண்டுபிடித்தல் (relation finding) முக்கியமான செயல்பாடாகும்.

ஆழமில்லாப் பகுப்பாய்விகள் இயற்கை மொழி முழுவதையும் கையாள வேண்டி இருப்பதால் அவை பெரிதாகவும் பெரும்பாலும் ஆயிரக்கணக்கான விதிகளையும் கொண்டிருக்கின்றன. இவ்விதிகளில் பல விதிவிலக்கானவை என்பதன் அடிப்படையில் இவை வலிமை குறைந்தவையாக இருக்கின்றன. ஆழமில்லாப் பகுப்பாய்விகள் இயந்திர கற்றல் அடிப்படையிலான தொழில்நுட்பங்களைப் (techniques) பயன்படுத்தித் தானியக்கமாக உருவாக்கப்படுகின்றன.

5.4. சொல் வகைப்பாட்டுக் குழுமங்களும் சொல் வகைப்பாடு அடையாளம் படுத்துகையும் (Parts of speeches and Part of Speech Tagging)

மொழி ஆய்வில் சொல் வகைப்பாட்டின் முக்கியத்துவம் என்னவென்றால் அது சொல்லைப்பற்றியும் அவற்றை அடுத்து வரும் சொற்கள் பற்றியும் செய்திகளை ஏராளமாகத் தருகின்றது. இவை முக்கிய வகைப்பாடுகளுக்கு (வினை, பெயர்) மட்டுமின்றி பல நுட்பமான வேறுபாடுகளுக்கும் உண்மையாகும். எடுத்துக்காட்டாக இந்த அடையாளக் குழுமங்கள் உடமை மாற்றுப் பெயர்களுக்கும் (என்னுடைய, உன்னுடைய, அவனுடைய, அவளுடைய, அவருடைய, அதனுடைய) உடைமையல்லா மாற்றுப் பெயர்களுக்கும் (நான், நீ, அவன், அவள், அது) வேறுபாட்டைக் காட்டுகின்றன. ஒரு சொல் உடமை மாற்றுப்பெயரா அல்லது உடமையலா மாற்றுப்பெயரா என்பதை அறிந்து கொண்டு எந்தச் சொற்கள் அவற்றின் அருகில் வர இயலும் என்பதைக் கூற இயலும். இதைப் பேச்சு தெரிந்துகொள்ளுதல் (speech recognition) என்பதற்கான மொழி மாதிரிகளில் (language models) பயன்படுத்தலாம். சொல்வகைப்பாடு, தகவல் மீட்டலுக்கான (Information reterival) வேர்ச்சொல்லைக் கண்டுபிடிப்பதற்குப் (stemming) பயன்படுகின்றது. ஏனென்றால் சொல் வகைப்பாட்டை (part of speech) அறிவதால் அவை எந்த உருபனியல் ஒட்டுகளை ஏற்கும் என்பதை அறியலாம். அவை பனுவலிருந்து பெயர்களையோ முக்கிய சொற்களையோ அறிய உதவுவதால் தகவல் மீட்டல் பயன்பாட்டிலும் பயன்படலாம். தானியங்கு சொல்வகைபாட்டு அடையாளப்படுத்திகள் சொல்பொருள் மயக்கம் நீக்கி வழியமைப்பு முறைகள் (Word Sense disambiguation algorithms), சொல் வகைப்பாட்டு அடையாளப்படுத்திகள் (Part-of-Speech Taggers (POS taggers)), வகுப்பு அடையாளப்படுத்திகள் போன்றன முன்னேறிய தானியக்க பேச்சு அறியும் (Automatic Speech Recognition (ASR)) மொழிமாதிரிகளில் பயன்படுத்தப்படுகிறன. தகவல் பிரித்தெடுப்பு பயன்பாடுகளிலும் **(Information Extraction Applications)** சொல் வகுப்பு அடையாளப்படுத்திகள் பயன்படுத்தப்படுகின்றன.

முன்னர் கூறியபடி விரிதரவுகளைச் சொல்வகைப்பாட்டுக்கு அடையாளப்படுத்தல் முக்கியமான செயல்பாடாகும். சொல்வகைப்பாடு அடையாளப்படுத்தப்படாத விரிதரவுகள் இயற்கை மொழி ஆய்வுக்குப் பயன்படுத்த முடியாதவை. குறிப்பாக இயந்திர மொழிபெயர்ப்புக்குச் சொல்வகைப்பாடு அடையாளப்படுத்தல் மிக முக்கியமானதாகும்..

5.4.1.சொல்வகைப்பாட்டு அடையாளக் குழுமங்களைக் கண்டுபிடிக்கும் முயற்சிகள்

சொல்வகை அடையாளப் படுத்தலில் முக்கியமான கட்டம் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களைக் கண்டுபிடிப்பதாகும். இக்குழும உறுப்பினர்களின் எண்ணிக்கை சொல்லவகைப்பாடு அடையாளப்படுத்தலின் நோக்கத்திற்கு ஏற்ப வேறுபடும். இது சிறிய எண்ணிக்கையிலான சொல்வகை அடையாளக் குழுமங்களிலிருந்து பெரிய எண்ணிக்கையிலான சொல்வகை அடையாளக் குழும் வரையிலான வேறுபாட்டைக் காட்டும். இது எண்ணிக்கை அடிப்படையிலும் நுணுக்கம் அடிப்படையிலும் அமைப்பு அடிப்படையிலும் (தட்டையான

அமைப்பைத் தரும் அடையாளக் குழுமம், படிநிலையான அமைப்பு தரும் அடையாளக் குழுமம்) என வேறுபடும்.

ஆரம்பக்கட்ட சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம் (1980-க்கு முன்)

1960-களிலும் 1970-களிலும் ஆங்கில மொழிக்கு வேண்டி சொல் வகைபாட்டு அடையாளக் குழுமங்களை உருவாக்கும் செயல்பாடு ஐக்கிய அமேரிக்காவில் தொடங்கப்பட்டுவிட்டது. ஆரம்பக் காலத்தின் முக்கியமான சொல்வகைபாட்டு அடையாளக் குழுமங்கள் க்ளைன்–சிம்மொன்ஸ் (Klein-simmons 1963) மற்றும் க்ரீன் ரூபின் (Green-Rubin 1971) என்பவர்களால் உருவாக்கப்பட்டது. மற்றொரு முயற்சி பென்சிலியா பல்கலைக்கழகத்தில் மேற்கொள்ளப்பட்டது.

தொடர்ச்சியான ஆங்கில சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் (1980க்குப் பிறகு)

1980-1990-களில் லங்காஸ்டர் பல்கலைக்கழகத்தில் ஆங்கிலத்திற்கு வேண்டி ஒரு தொடர்ச்சியான சொல்வகைபாட்டு குழுமங்கள் CLAWS என்ற சொல் வகைப்பாட்டை அடையாளப்படுத்தும் மென்பொருளைப் பயன்படுத்தி உருவாக்கப்பட்டது. CLAWS என்ற சொல்வகைப்பாட்டு அடையாளக் குழுமம் LOB விரிதரவுக்குப் பயன்படுத்தப்பட்டது. பின்னர் CLAWS-இன் C5 மற்றும் C7 சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் Bஈஊ விரிதரவை அடையாளப்படுத்த பயன்ப்படுத்தப்பட்டது. இதைத் தொடர்ந்து பல ஆங்கில சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் பயன்படுத்தப்பட்டன. அவையாவன:

1. The TOSCA Scheme
2. The ICE Tagset
3. The PENN Tagset
4. The LUND Tagset
5. The Eng CG Tagset

ஆங்கிலத்தில் புகழ்பெற்ற POS Taggers Brown Corpus-இல் பயன்படுத்தப்பட்ட 87 அடையாளக் குழுமங்களிலிருந்து (tagsets) உருவாக்கப்பட்டவை. மூன்று மிகப் பொதுவாகப் பயன்படுத்தப்படுகின்றவை:

1. 45 – Tag penn treebank tagset
2. The Mediurn size 61 tag C5 Tagset used by Lancaster VCREL Projects CLAWS (Constituent Liksihood Automatic word – tagging system) tagger to tag British National Corpus (BNC)
3. The Larger 146 – tag C7 tagset

Ex

Penn tree bank version of brown corpus ASCII flat file

Tag represented after each eord following a slash The/ DT grand /JJ/ jury/NN Commented/ VBD on/ IN a/DT number/ NN of?IN others/ JJ topies/ NNs/

DT - Determiners

JJ - Adjective

NN - Noun, sing, or mass

VBD - Verb, Past tense

IN - Prepossition

NNS - Noun, Plural

5.4.2.சொல்வகைப்பாட்டு அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பு (The EAGLES – Guidelines)

ஆங்கில தவிர பிற மொழிகளுக்குச் சொல் வகைப்பாடு அடையாளம் செய்யும் போது சொல்வகைப்பாட்டு அடையாளக் குழுமங்களைத் தரப்படுத்தும் தேவை வெளிப்பட்டது. சொல் வகைப்பாட்டு அடையாளக் குழுமங்களின் மிகச் சமீபக்காலத்திய தரமான ஒன்று EAGLES– வழிகாட்டல்களாகும் (Leeach and Wilson). இதனுடைய நோக்கம் பலவிதமான திட்டங்களுக்கும் பல மொழிகளுக்கும் சொல்வகைப்பாட்டு அடையாக் குழுமங்களைத் தரப்படுத்துதவதாகும். EAGLES வழிகாட்டல்கள் சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் உள்ளடக்கும் அல்லது உள்ளடக்க வேண்டிய ஒரு குழுமப் பண்புக்கூறுகளைத் தருகிறது. இதனுடன் இந்த எல்லாப் பண்புக்கூறுகளையும் குறியாக்கம் செய்து இடைநிலைப்பட்ட குறியாக்கத்தின் தேவை என்னவென்றால் EAGLES வழிகாட்டல்களின் உதவியுடன் உருவாக்கப்பட்ட இரண்டு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களுக்கு இடையில் மொழிபெயர்ப்பை அனுமதிப்பாதாகும். இதனால் அவற்றிற்கிடையில் பொருத்ததை உறுதி செய்வதாகும். உருப-தொடரியல் அடையாளப்படுத்தலுக்கு EAGLES வழிகாட்டல்களில் மூன்று கட்டங்களில் அமைப்பாக்கம் செய்யப்பட்டுள்ளது.

1. எவை எவற்றைக் கட்டாயமாகக் கருதப்படுகிறது.

2. எவை பரிந்துரைக்கப்படுகின்றது.

3. எவை உருப-தொடரியல் அடையாளப்படுத்தலுக்கு மொழிசார் சிறப்புள்ளது அல்லது சிறப்பற்றது என அமையும் பண்புகளுக்கு விருப்பு நீட்சிகள்

ஒவ்வொரு மட்டத்திலும் அடையாளங்கள் உருப-தொடரியல் சின்னம்-மதிப்பிணைகளாக வரையறை விளக்கம் செய்யப்படுகிறது. எடுத்துக்காட்டாக 'பால்' என்ற சின்னத்திற்கு EAGLES-

இன் பரிந்துரைகளில் ஆண், பெண் அல்லது பால் சாராதவைகள் என்ற மதிப்புகள் இருக்கும். சின்னம்-மதிப்பிணைகள் படிநிலைகளாக அமைக்கப்படவோ அமைக்கப்படாமலோ இருக்கலாம்.

EAGLES வழிகாட்டல்களால் எந்தச் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களுக்கும் கட்டாயமாகப் பதிமூன்று பரிந்துரைக்கப்படுகிறது.

1. Noun (பெயர்)
2. Verb (வினை)
3. Adjective (பெயரடை)
4. Pronoun/determiner (பதிலீடுபெயர்)
5. Article (சுட்டிடைச்சொல்)
6. Adverb (வினையடை)
7. Ad position (உருபு)
8. Conjunction (இணைப்பான்)
9. Numeral (எண்)
10. Interjection (வியப்பிடைச்சொல்)
11. Unassigned/Unique (இணையற்ற/தனிப்பட்ட)
12. Residual (மீதியான)
13. Punctuation (நிறுத்தற்குறிகள்)

பரிந்துரைக்கப்பட்ட மற்றும் இருப்புச் சின்னங்கள் இந்த முக்கியமான சொல் வகைப்பாடுகளால் ஒழுங்குபடுத்தப்பட்டுள்ளது; சொல் வகைகளைக் கடந்து பொருத்தமுறத் தேவையில்லை.

EAGLES வழிகாட்டல்கள் சொல்வகைப்பாட்டு அடையாளக் குழுமத்தைத் திட்டமிடுபவரின் சுதந்திரத்தைக் கட்டுப்படுத்தாமல் அவர் குறிப்பிட்ட எண்ணும் எல்லா அடிப்படையான செய்திகளையும் உள்ளடக்கிய அடையாளக் குழுமங்களை EAGLES வழிகாகட்டல்களில் எளிதில் பொருந்தாத பண்புக் கூறுகளை உள்ளடக்கலாம். இருப்பினும் இவைகள் மற்றொரு அடையாளக் குழுமத்திற்கு மொழிபெயர்க்க இயலாத பண்புக்கூறுகளாகும். ஏனென்றால் எந்த அடையாளக் குழுமமும் அவற்றை உள்ளடக்கவில்லை. எனவே அவற்றை EAGLES இடைப்பட்ட அடையாளக் குழுமத்தில் அடக்குவது தேவையில்லாததாகும். இந்தச் சட்டகத்தின் திண்மையையும் திரும்பப் பயன்படுத்தலையும் ஊக்குவிக்கிறது. மற்றும் சுழலான திரும்பக் கண்டுபிடித்தலை ஊக்கமிழக்கச் செய்கிறது.

EAGLES வழிகாட்டலின் முக்கியமான குறை இது. உலக மொழிகளின் மிகக் குறுகிய பகுதியைதான் அடக்குகிறது. ஐரோப்பிய ஒருங்கிணைப்பின் திட்டமாக இது ஆங்கிலம், டச்சு ஜொ;மனி, டேனிஷ், பிரெஞ்சு, ஸ்பேனிஷ், போர்த்துகீசு, கிரீக், இத்தலியன் என்ற

வகைபாட்டியல் அடிப்படையில் ஒற்றுமையுள்ள ஒன்பது மொழிகளைத்தான் உள்ளடக்குகிறது. எனவே இவை அல்லாத பிற மொழிகளுக்கு EAGLES வழிகாட்டல்களின் பயன்பாடு பொருத்தமாகவோ பொருத்தமற்றோ இருக்கலாம்.

EAGLES **வழிகாட்டல்களின் அடிப்படையில் உருவாக்கப்பட்ட சில சமீபக் காலத்திய சொல் வகைப்பாட்டு அடையாக் குழுமங்கள்**

EAGLES வழிகாட்டலின் வெளியீட்டிற்குப் பிறகு அவற்றைப் பயன்படுத்தி பலவிதமான மொழிகளுக்குச் சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் உருவாக்கப்பட்டன. அவை பின்வருமாறு:

1. The MULTEXT[20] Project
2. The GRACE Project
3. The CRATER Project

EAGLES வழிகாட்டல்களால் அடக்கப்படாத மொழிகளின் சமீப காலத்திய சொல்வகைப்பாட்டு அடையாளக் குழுமங்களில் பின்வருமாறு அடங்கும்:

1. Tagset Design for Arabic
2. Tagset Design for Chinese
3. Tagset Design for Korean
4. Tagset Design in the Paninjan Tradition

5.4.3. இந்தியமொழிகளின் மையநிறுவனத்தின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள்

இந்திய மொழிகளுக்கு விரிதரவுகள் உருவாக்குவதில் முன்னோடியாக விளங்கிய இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்திய மொழிகளுக்கு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களை உருவாக்கி விரிதரவுகளை அடையாளப்படுத்தியுள்ளது. 2006-இல் இந்தி விரிதரவுக்குச் சொல்வகைப்பாட்டு அடையாளக் குழுமமாக முடிவுசெய்யப்பட்ட சொல்வகைப்பாட்டு குழுமத்தின் பட்டியல் கீழே தரப்பட்டுள்ளது.

CIIL Hindi Corpora Tag set Finalized, August 29, 2006

NOUN

1.	NN	Common Noun
2.	NNP	proper Noun (place, person, etc)
3.	NC	Noun compound
4.	NAB	Noun Abstract
5.	CRD	Cardinal number
6.	ORD	Ordinal number

		PRONOUN	
7.	PRP		Personal Pronoun
8.	PRI		Indefinite pronoun
9.	PRR		Pronoun Reflexive
10.	PRL		Relative Pronoun
11.	PDP		Demonstrative pronoun
		VERB	
12.	VF		Verb Finite Main
13.	VNF		Verb non-finite adjectival
			Verb non-finite Adverbial
14.	VAX		Verb Causative
15.	VNN		Gerund/Verb non-finite nominal
16.	VINF		Verb infinitive
17.	VCC		Verb Causative
18.	VCD		Verb Double Causative
		ADJECTIVE	
19.	ADD		Adjective-declinable
20.	ADI		Adjective-indeclinable
		ADVERB	
21.	ADV		Adverb
22.	ADL		Adverb of location
		OTHERS	
23.	QWO		Question Words
24.	PPO		Post Position
25.	INT		Intensifier
26.	QTE		Quantifier
27.	CNJ		Conjunction
28.	INJ		Interjection
29.	IND		Indeclinable

30.	QOT	Quotative
31.	PRT	Particle
32.	NEG	Negative
33.	RDP	Reduplication
34.	FWD	Loan Word (non-notarized words)
35.	IMD	Idiom
36.	PRO	Proverb

5.4.4.எயுகேபிசி ஆய்வுமையத்தின் சொல்வகைப்பாட்டு அடையாளக் குழுமம்

AUKBC நிறுவனம் தமிழுக்காக ஒரு சொல்வகைப்பாட்டு அடையாளக் குழுமத்தை உருவாக்கியுள்ளது. இது எல்லா இலக்கண மற்றும் சொல்சார் உறுப்புகளையும் உள்ளடக்கிய முழுமையான சொல்வகைப்பாட்டு அடையாளக் குழுமம் ஆகும். இது தமிழ்ப் பல்கலைக்கழக மொழியியல் வல்லுநர்கள் உதவியால் AUKBC யின் NLP குழுவினரால் 2001-இல் திட்டமிடப்பட்ட சொல்வகைப்பாட்டு அடையாளக் குழுமமாகும். பின்வரும் பட்டியல் தமிழின் சொல்வகைப்பாட்டு அடையாளக் குழுமமாகும்.

TAMIL TAGSET

No	Description	TAG	Examples	Suffixes
1.	Noun Singular	N. Sg	mar am (tree)	
2.	Noun Plural	N.Pl	marafkal Y(trees),	
3.	Pronoun	Pr.Obl.Sg.	eVnY (mine)	
4.	Pronoun Oblige From	Pr.Obl.Sg.	eVnY (mine)	
5.	Pronoun Plural	Pro.Pl.	avarkalY (they)	
6.	Pronoun Plural Oblige	Pro.Obl.Pl.	eVfkalY (ours), ufkaly (yours)	
7.	#Euphonic increment	Eu	Pacuvin Y (of the cow)	
8.	Acusative	Acc	Avan Y Ei (him accusative)	E
9.	Dative	Dat	avan Y ukku (to)	Ukku, ku, kku
10.	Instrumental	Inst	cAviyAL (by the	Al

			key)	
11.	Sociative	Soc	avan yotu (with him), Avan Yutan Y (with him)	
12.	Locative	Loc	pUfkAvil (in the park)	il
13.	Ablative	Abl	Urilirunwu (from city/town	ilirunwu
14.	Benefactive	Ben	aval YukkAka (for her	ukkAka
15.	Genetive	Gen	Pacuvin (of the cow), utEya (of), avan Yawu (his)	In Y, uEya, awu
16.	Vocative	Voc		Ata, ivalYe
17.	Clitics	Cl		A,o,e,um,wAnY
18.	Verb	V	Pati (read)	
19.	#Transitive Verb	VT	wiruppu(turn)	
20.	#Intransitive Verb	VI	Wirumpu (turn sth)	
21.	#Causative	VC	Werivi(inform)	
22.	Infinitive	Inf	Ceyya (to do)	a
23.	Singular Negative (imperative)	Neg	ceyAwe (you)don't do)	Awe
24.	Plural Imperative	V.Pl	ceyyufkalY ((you) do-hon or plu)	Ufkal Y
25.	Plural Negative (imperative)	Pl.Neg	Ceyyufkal Y ((you) do-hon or plu)	A wirkal Y
26.	Imperative + suffixes	Imp	vAta, ammA	

27.	Present Tense	Pre		kirY,kkirY.
28.	Past Tense	Pst		W,nw,in,ww,I,t,rY
29.	Future Tense	Fut		V,p,pp,um
30.	First Person Singular	IS		enY
31.	First Peprson Plural	IPI		om
32.	Second Person Singular	2S		Ay
33.	Second Person Plural/Honorific	2PI		Ir,Irkal Y
34.	Third Person Masculine Singular	3MS		AnY
35.	Third Person Feminine Singular	3FS		AIY
36.	Third Person Honorific Singular	3FS		Ar,ar,anYar
37.	Third Person Singular Neuter (non-human)	3HPI		ArkalY
38.	Third Person Honorific Plural	3HPI		ArkalY
39.	Third Person Plural Neuter (non-human)	3NPL		anYa,um
40.	Third Person Neuter Future Negative	HumFutNeg	mAttenY((I) won't)	mAtt
41.	Future Negative	HumFut	varavollE (didn't	illE

	Human	Neg	come)	
42.	Present/Past Negative (non-future)	Nfut Neg	varavillE (didn't come)	illE
43.	#Optative	Opt	vAIYYka(long live)	Ka,kka
44.	Verbal Participle	Vbp	ceVywu (done)	wu,nwu,etc.
45.	Verbal Participle Negative	Vbp Neg	ceVyyAmal (not done)	Amal,Awu
46.	Conditional Positive	Cnd	ceVywAl(if done)	Ai
47.	Conditional Negative	Cnd Neg	ceVyy AvittAl (if not done)	AvittAl
48.	Adjectival Participle Present/Past	Adjp	ceVywa (do pre/pst sth), ceVykiYa (do pre/pst.th)	wa, kir Ya
49.	Adjectival Participle Future	AdjPfut	ceVyyum (do future sth)	um
50.	Adjectival participle Negative	AdjpNeg	ceVyAwa (don't do sth)	Awa
51.	Verbal Noun Untensed	VbnUnt	ceVyyal (doing), ceVywal (doing), patikkal (reading), patiwwal (Reading), ceVykal (doing), patikkai (reading)	al,wal,kkal,wwal,kai,kkai
52.	Verbal Noun Tensed Positive	Vbn	ceVywawu (have done)	awu

53.	Verbal Noun Tensed Negative	VbnNeg	ceVyAwawu (have not done)	Awawu
54.	Participal Noun	PpIn	ceVywavanY (one who did)	anY,alY, wu, ar, arkal Y, ai
55.	Adjectival Noun	AdjN	Nallavan Y (good man)	anY, alY, wu, ar, arkal Y, ai
56.	Selective (Concessive?)	Sel	avanyAvaWu (at least him)	Avawu
57.	Oridinal Suffix	Ord	muwalAvawu (first)	Avawu
58.	Reportative Suffix	Rep		Am
59.	Complementizer	Cmp		eVnYrYu, eVnYpawu, Aka
60.	Auxiliary	Aux		
61.	Conjuntion	CnJ		AnYAI,illEyeVnYrYAI,
62.	Adjectives	Adj	AlYYakAnYa (beautiful-with adj suffix)	AnYa
63.	Adverbs	Adv	alYYakAka (beautiful-with adj suffix)	AnYa
64.	Postpositions	PSP	Mel (above), kllYY (below)	
65.	Quantifiers	Q	oVnYrYu (one), cila (few), koFcam (some)	
66.	Determiners	Det	Anwa (that), inwa (this)	
67.	Wh Words	INT	eVnYnYa (what), eVvai (which),	

			eVppati (how)	
68.	Unknown	UNK		

மைய அரசின் நிதி உதவியுடன் நடைபெற்ற இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்புக்காகப் பயன்படுத்தப்பட்ட WXT என்ற ரோமன் எழுத்துமுறையைப் பயன்படுத்தி தமிழ் எடுத்துக்காட்டுகள் தரப்பட்டுள்ளன. இதை படிப்பது கடினம். இயந்திரம் புரிந்துகொள்வதற்காக இம்முறை பயன்படுத்தப்பட்டது.

5.4.5. ஹைதராபாத் IIIT-ஆல் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமம்

இந்திய மொழிகளுக்கிடையிலான மொழிபெயர்ப்பிலும் ஆங்கில மொழியிலிருந்து இந்திய இயந்திர மொழிபெயர்ப்பிலும் ஈடுபட்டுள்ள ஹைதராபாத்தில் உள்ள IIIT நிறுவனம் இந்தியாவில் இயற்கைமொழி ஆய்வுக்கு முன்னோடியாக உள்ளது. இந்நிறுவனம் உருவாக்கிய இந்திய மொழிகளுக்கான (குறிப்பாக இந்திக்கான) சொல்வகைப்பாட்டு அடையாளக் குழுமம் பரவலாகப் பயன்படுத்தப்படுகிறது. இதன் பட்டியல் கீழே தரப்பட்டுள்ளது.

NN (Common noun)

NNP (Proper Noun)

*C (for all compounds)

QC (Cardinal No.)

QO (Ordinal No.)

PRP (Pronoun)

VF (Verb Auxiliary)

VNF (Verb Non-Finite adverbial and adjectival)

VAUX (Verb Auxiliary)

VNN (Gerund/Verb non-finite nominal)

VINF (Verb Infinitive)

JJ (Adjectives)

RB (Adverbs)

PSP (Postposition)

QF (Question words)

WQ (Question words)

CC (Conjunct)

INT (Intensifier)

NEG (Negative)

INJ (Interjection)

RP (Particle)

CL (Classifier)

SVM (Special)

AUKBC, CIIL, IIIT-H **இந்நிறுவனங்களின சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு**

AUKBC, CIIL, IIIT-H இந்நிறுவனங்களில் உருவாக்கப்பட்ட சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீட்டுப் பட்டியல்கள் கீழே தரப்பட்டுள்ளன. (இப்பட்டியல் AUKBC நிறுவன அறிவியலாளர் ஷோபா என்பவரால் உருவாக்கப்பட்டது).

NOUNS

AU_KBC	CIIL	IIT-H
	NNP (Proper Noun)	NNP (Proper Nopun)
	NC (Noun Compound)	*C (for all compounds)
	NAB (abstract noun)	
	CRD (Cardinal No.)	QC (Cardinal No.)
Ord (ordinal suffix)	ORD (Ordinal No.)	QO (Ordinal No.)
N.Sg (Noun Sg)		
N.Pl (Noun Plural		
	PRP (Indefinite Pronoun)	
	PRI (Indefinite Pronoun)	
	PRR (Reflexive Pronoun)	
	PDP (Demonstrative)	
Pr.Sg. (Pronoun Sg)		
Pr.Pl. (Pronoun Plural)		
Pr.Obl.Sg. (Pronoun Oblique Sg.)		
Pr. Obl.Pl (Pronoun Oblique		

Plural)		
V(Verb)	VF (Verb Finite Main)	VF (Verb Finite Main)
	VNF (Verb Non-Finite adverbial and adjectival	VNF (Verb Non-Finite adverbial and adjectival
	VAX (Verb Auxiliary)	VAUS (Verb Auxiliary)
VbnUnt (Verbal Noun Untensed Vbn (Vbn Tensed Positive VbnNeg (Vbn Tensed Negative	VNN (gerund/Verb non-finite nominal)	VNN (Gerund/Verb non-finite nominal)
Inf (Infinitive)	VINF (Verb Infinitive)	VINF (Verb Infinitieve)
VC (Verb Transitive)		
VT (Verb Transitive)		
VI (Verb Intransitive)		

ADJECTIVES

AU-KBC	CIIL	IIIT_H
Adj (Adjectives)		JJ (Adjectives)
	ADD (Adjective Declinable)	
	ADI (Adjective Indeclinable)	

ADVERBS

AU-KBC	CIIL	IIIT-H
Adv (Adverbs)	ADV (Adverbs)	RB (Adverbs)
	ADL (Adverb of location)	

OTHERS

AU-KBC	CIIL	IIIT-H
PSP (Postposition)	PPO (Postposition)	PSP (Postposition)
Q (Quantifiers)	QTF (Quantifiers)	QF (Quantifiers)
Cnj (Conjunction)	CNJ (Conjunction)	CC (Conjunct)
	INT (Intersifier0	INT (Intensifier)

	NEG (Negative)	NEG (Negative)
	INJ (Interjection)	INJ (Interjection)
	PRT (Particle)	RP (Particle)

OTHERS

CIIL
IND (Indeclinable)
QOT (Quotative)
RDP (Reduplication)
LWD (Loan Word)
IDM (Idion)
PRO (Proverb)

5.4.6.அமிர்தா பல்கலைக்கழத்தின் சொல்வகைப்பாட்டு அடையாளக்குழுமம்

அமிர்தா பல்கலைக்கழகம் தமிழ் தழுவிய இயற்கைமொழியியல் ஆய்வுக்காகப் பின்வரும் அடையாளகுழுமத்தை நிறுவியுள்ளது.

S.N	TAG	DESCRIPTION		S.N	TAG	DESCRIPTION
1	<NN>	NOUN		17	<VINT>	VERB INFINITE
2	<NNC>	COMPOUND NOUN		18	<CNJ>	CONJUNCTION
3	<NNP>	PROPER NOUN		19	<CVB>	CONDITIONAL VERB
4	<NNPC>	COMPOUND PROPER NOUN		20	<QW>	QUESTION WORDS
5	<ORD>	ORDINALS		21	<COM>	COMPLEMENTIZER
6	<CRD>	CARDINALS		22	<NNQ>	QUANTITY NOUN
7	<PRP>	PERSONAL PRONOUN		23	<QTF>	QUANTIFIERS
8	<PRIN>	PRONOUN INTROGATIVE		24	<PPO>	POSTPOSITIONS
9	<PRID>	PRONOUN INDEFINITE		25	<DET>	DETERMINERS
10	<ADJ>	ADJECTIVE		26	<INT>	INTENSIFIER
11	<ADV>	ADVERB		27	<ECH>	ECHO WORDS
12	<VNAJ>	VERB NON FINITE ADJECTIVE		28	<EMP>	EMPHASIS
13	<VNAV>	VERB NON FINITE ADVERB		29	<COMM>	COMMA
14	<VBG>	VERBAL GERUND		30	<DOT>	DOT
15	<VF>	VERB FINITE		31	<QM>	QUESTION MARK
16	<VAX>	VERB AUXILARY		32	<RDW>	REDUPLICATION WORDS

5.4.7. தமிழுக்கான பிற சொல்வகைப்பாட்டு அடையாளப்படுத்திகள்

தமிழுக்கு வேண்டி பல சொல்வகை அடையாளப்படுத்திகள் உருவாக்கப்பட்டுள்ளன. அவற்றில் சில முக்கியமானவை கீழே தரப்பட்டுள்ளன.

5.4.7.1 வாசு அரங்கநாதனின் டேக் தமிழ்

வாசு அரங்கநாதனால் உருவாக்கப்பட்ட டேக் தமிழ்ச் (Vasu Ranganathan's Tag Tamil) சொல்-ஒலியனியல் (Lexical-phonology) அணுகுமுறை அடிப்படையில் உருவாக்கப்பட்டது. டேக் தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வை உருபன்களின் வருகை முறையை வரிசை அட்டவணையைப் பயன்படுத்தி செய்கின்றது. டேக் தமிழ் சொல் வகை அடையாளப்படுத்தலையும் உருவாக்கலையும் செய்கிறது.

5.4.7.2 கணேசனின் சொல்வகைப்பாட்டு அடையாளப்படுத்தி

கணேசன் தமிழுக்கு வேண்டி ஒரு சொல் வகை அடையாளப்படுத்தியை உருவாக்கியுள்ளார். அவருடைய அடையாளப்படுத்தி இந்திய மொழிகளின் மைய நிறுவன தமிழ் விரிதரவில் நன்கு வேலை செய்கின்றது. பிற தரவுத்தொகுதிகளின் இதன் திறன் மதிப்பீடு செய்யப்பட வேண்டும். இவர் தமிழுக்கு வேண்டி விரிவான சொல்வகை அடையாளப்படுத்திகளின் குழுமத்தை வைத்திருக்கின்றார். இவர் இந்திய மொழிகளின் மைய நிறுவன தரவுத்தொகுதியின் ஒரு பகுதியை அகராதியாலும் உருபனியல் பகுப்பாய்வியாலும் சொல் வகைக்காக அடையாளப்படுத்தி உள்ளார். பின்னர் இவை அடையாளப்படுத்தப்பட்ட விரிதரவை மனிதசக்தியால் திருத்தி மீதியுள்ள பெருந்தரவுக்கு பயிற்சி தந்துள்ளார். இவருடைய சொல்வகை அடையாளப்படுத்தியில் சொல்வகை அடையாளப்படுத்தல் ஒவ்வொரு உருபன்களுக்கும் தரப்படுகின்றது.

5.4.7.3 RCILTS-இன் தமிழ் கதம்பம்

கதம்பம் ஆவணத்தில் தரப்பட்டுள்ள சொல்லுக்குச் சொல்வகை அடையாளத்தைத் தருகிறது. இது தமிழ்மொழியியல் அடிப்படையிலான ஊக விதிகளைப் பயன்படுத்துகின்றது. அகராதியையோ உருபனியல் பகுப்பாய்வியையோ பயன்படுத்துவதில்லை. இது பெரிய ஆவணங்களில் 80% திறனைத் தருகிறது. இது பன்னிரண்டு ஊக விதிகளைப் பயன்படுத்துகிறது. இது இட-எண்-பால், காலம் மற்றும் வேற்றுமை குறியீடுகளின் அடிப்படையில் சொல்வகை அடையாளங்களை அடையாளம் காண்கிறது. தனியாக நிற்கும் சொற்கள் சொல்வகை அடையாளப்படுத்தியில் சேகரிக்கப்பட்ட பட்டியலில் இருந்து பரிசோதிக்கப்படுகிறது. இது அறியப்படாத சொற்களுக்கு (unknown words) நிரப்பு விதியைப் பயன்படுத்துகிறது. இது bபைசயஅ மாதிரிகளைப் பயன்படுத்தி முன்வரும் சொல்லின் சொல்வகைப்பாட்டைக் கொண்டு அறியப்படாத சொல்லை அடையாளம் காண்கிறது.

5.4.7.4. பாஸ்கரன் மற்றும் பிறரின் இந்திய மொழிகளுக்கான பொது சொல்வகைப்பாட்டு அடையாளக்குழுமச் சட்டகம்

பாஸ்கரன் மற்றும் பிற மொழியியல் வல்லுனர்களும் (Baskaran et al 2007, 2008) கணினி மொழியியல் வல்லுனர்களும் இணைந்து பெங்களூரிலுள்ள மைக்ரோசாஃப்ட் ஆய்வு நிறுவனத்தின் முயற்சியால் இந்திய மொழிகளுக்கான பொது சொல்வகை அடையாளக் குழுமச் சட்டகத்தை உருவாக்கினர். இது படிநிலை அமைப்பு அடிப்படையிலான சொல்வகை குழுமம் ஆகும். இந்த அடையாளக் குழுமம் வகைப்பாடுகள், வகைகள், பண்புகள் என பிரிக்கப்பட்டு பட்டியலிடப்பட்டுள்ளன. பின்வரும் அட்டவணை இந்திய மொழிகளுக்கான பொது சொல்வகை அடையாளக்குழுமச் சட்டகம் ஆகும்.

Level-0	Level-1	Level 2			
Lexical categories	**Types**	**Attributes**	**Values**	**Types Applicable**	**Remarks**
[N] Nouns	[1] Common [2] Proper [3] Verbal [4] Spatial/ temporal Adjectival nouns such as நல்லவன் will be tagged as common nouns	Gender	[1] Masculine [2] Feminine [3] Neuter	[1] Common	
		Number	[1] Singular [2] Plural [3] Plural/ honourific	[1] Common [2] Proper	
		Case*	[1] Direct [2] Oblique	[1] Common	
		Case-markers	[2] Accusative (*E*) [3] Instrumental (*Al*) [4] Dative (*kku* or *ukku*) [5] Genitive (*in* | *utEya* | *athu*) [6] Sociative (*otu* | *utan*) [7] Locative (*il* | *itam*)	[1] Common [2] Proper [3] Verbal [4] Spatial/ termporal	

			[8] Ablative (*irunthu*) [9] Benefactive (*ukkAka*)		
		Negative	[1] Yes	[1] Common [2] Proper [4] Spatial/ termporal	
		Emphatic*	[1] Yes	[1] Common [2] Proper [4] Spatial/ termporal	
		Definitenes s*	[1] Yes	[1] Common [2] Proper [4] Spatial/ termporal	
[V] Verbs	[1] Finite [2] Auxiliary [3] Infinitive [4] Nonfinite Non-finites forms will betagged as participles	Gender	[1] Masculine [2] Feminine [3] Neuter	[1] Finite [2] Auxiliary	
		Number	[1] Singular [2] Plural [3] Plural/ honourific	[1] Finite [2] Auxiliary	
		Person	[1] First [2] Second [3] Third	[1] Finite [2] Auxiliary	
		Tense	[1] Past [2] Present [3] Future	[1] Finite [2] Auxiliary	
		Aspect*	[1] Simple [2] Continuous	[2] Auxiliary	Dravidian has separate

			(progressive) [3] Perfect		auxiliaries for aspect and mood. However, many a times these are agglutinated with the main verb. Here it is recommended that the word be segmented first before being tagged.
		Mood*	[1] Declarative [2] Subjunctive [3] Conditional [4] Imperative [5] Presumptive	[2] Auxiliary	
		Negative	[1] Yes	[1] Finite [2] Auxiliary [3] Nonfinite	
		Emphatic*	[1] Yes	[1] Finite [2] Auxiliary [3] Infinitive	
[PR] Pronouns	[1] Personal [2] Possessive [3] Reflexives [4] Reciprocals [5] Spatial/ temporal	Gender	[1] Masculine [2] Feminine [3] Neuter	[1] Personal [2] Possessive	
		Number	[1] Singular [2] Plural [3] Plural/honourific	[1] Personal [2] Possessive	

		Person	[1] First [2] Second [3] Third	[1] Personal [2] Possessive	
		Case	[1] Direct [2] Oblique	[1] Personal [2] Possessive	
		Case-markers	[2] Accusative (*E*) [3] Instrumental (*Al*) [4] Dative (*kku* or *ukku*) [5] Genitive (*in* | *utEya* | *athu*) [6] Sociative (*otu* | *utan*) [7] Locative (*il* | *itam*) [8] Ablative (*irunthu*) [9] Benefactive (*ukkAka*)	[1] Personal [2] Possessive [3] Reflexive [5] Spatial/ temporal	
		Referential	[1] Absolute [2] Relative [3] Wh-words	[1] Personal [2] Possessive	
		Negative	[1] Yes	[1] Personal [2] Possessive [5] Spatial/ temporal	
		Emphatic*	[1] Yes	[1] Personal	

				[2] Possessive [5] Spatial/ temporal	
		Definitenes s*	[1] Yes	[1] Personal [2] Possessive [5] Spatial/ temporal	
		Incl/ Excl.*	[1] Yes	[1] Personal	
[NM] Nominal modifiers	[1] Adjectives [2] Demonstrativ es [3] Quantifiers	Referential	[1] Absolute [2] Relative [3] Wh-words	[2] Demonstrativ es	
		Numeral	[1] Non-numeral [2] Cardinal number [3] Ordinal number	[3] Quantifier	
		Negative	[1] Yes	[1] Adjectives [3] Quantifiers	
		Emphatic*	[1] Yes	[2] Demonstrativ es [3] Quantifiers	
[RB] Adverbs	-	Negative	[1] Yes	-	
[PL] Participles	[1] Relative [2] Verbal [3] Nominal [4]	Gender	[1] Masculine [2] Feminine [3] Neuter	[3] Nominal	*ceykiRavan*, *ceythathu*, *ceythavarka L*, *ceypavar*
		Number	[1] Singular	[3] Nominal	

	Conditional		[2] Plural [3] Plural/ honourific		Should these words be tagged as nominal participles?
		Case-markers	[2] Accusative (*E*) [3] Instrumental (*Al*) [4] Dative (*kku* or *ukku*) [5] Genitive (*in* | *utEya* | *athu*) [6] Sociative (*otu* | *utan*) [7] Locative (*il* | *itam*) [8] Ablative (*irunthu*) [9] Benefactive (*ukkAka*)	[3] Nominal	
		Tense	[1] Past [2] Present [3] Future	[1] Relative [3] Nominal	
		Aspect*	[1] Simple [2] Continuous [3] Perfect	[3] Nominal	
		Negative	[1] Yes	[2] Relative [3] Nominal [4]	

				Conditional	
		Emphatic*	[1] Yes	[3] Nominal [4] Conditional	
		Realis*	[1] Realis [2] Irrealis	[4] Conditional	Marks if the event is real or hypothesised one. Eg. வந்ததால் *vs.* வந்தால்
[PP] Postpositions	-	Emphatic*	[1] Yes	-	
[RP] Particles	[1] Coordinating [2] Subordinating [3] Interjection [4] Others	-	-	-	
[PU] Punctuation	-	-	-	-	
[RD] Residual	-	-	-	-	

5.4.7.5. இந்திய மொழிகளுக்கான பிஸ் சொல்வகைப்பாட்டு அடையாளக் குழுமம் (Bis tagset for Indian Languages)

இந்திய அரசின் நிதி நல்கையின் கீழ் நடைபெற்ற இயந்திர மொழி பெயர்பு, தரவுத்தொகுதி உருவாக்கம் என்ற ஆய்வுத்திட்டங்களுக்குத் தேவையான சொல்வகை அடையாளப் படுத்தப்பட்ட உரைகளை உருவாக்க இத்திட்டங்களில் பங்கேற்ற நிறுவனங்களின் கூட்டு முயற்சியால் Bis (Bureau of Indian Standards) tagset for Indian Languages உருவாக்கபட்டு பயன்பாட்டில் உள்ளது. இது ஒரு படிநிலை சொல் அடையாளக் குழுமமாகு. இவ்வடையாளக் குழுமம் கீழே பட்டியலிடப்பட்டுள்ளது.

S.No.	English		தமிழ் நிகரன்கள்	எடுத்துக்காட்டுகள்
1	Noun	N	பெயர்	மாம்பழம்

	common	N_NN	பொதுப்பெயர்	மாம்பழம்
	Proper	N_NNP	இயற்பெயர்	கண்ணன்
	Verbal		தொழிபெயர்	படித்தல்
	Nloc	N_NST	இடகாலப்பெயர்	பின்னால், பிறகு
2	Pronoun	*PRP*	மாற்றுப் பெயர்	நான், நீ,அவன்
	Personal	PR_PRP	மூவிடப்பெயர்	நான், நீ, அவன்
	Reflexive	PR_PRF	தற்சிட்டுப்பெயர்	தான்
	Reciprocal	PR_PRC	பரிமாற்றப்பெயர்	ஒருவரையொருவர்
	Relative	PR_PRL	சார்பு மாற்றுப்பெயர்	
	Wh-words	PR_PRQ	வினா மாற்றுப்பெயர்	யார்
3	Demonstrative	DM	சுட்டு	அந்த
	Deictic	DM_DMD	சுட்டு	அந்த
	Relative	DM_DMR	சார்பு	
	Wh-words	DM_DMQ	வினாச்சொல்	எது
4	Verb	*V*	வினை	பாடு, செய்
	Auxiliary Verb	V_VAUX	துணைவினை	இரு, கொண்டிரு
	Main Verb	V_VM	முதன்மைவினை	நட, ஓடு
	Finite	V_VM_VF	முற்றுவினை	ஓடினான்
	Infinitive	V_VM_VNF	வினை எச்சம்	ஓட, பாட
	Gerund	V_VM_VNG	வினைப் பெயர்	ஓடுதல், பாடுதல்
	Non-Finite	V_VM_VNF	எச்சவினை	வந்து, வந்த
5	Adjective	JJ	பெயரடை	நல்ல, பெரிய
6	Adverb	RB	வினையடை	மெல்ல, விரைவாக
7	Post Position	PSP	பின்னுருபு	கொண்டு, விட, காட்டிலும்
8	Conjunction		இணைப்புக்கிளவி	மற்றும்
	Co-ordinator	CCD	சமநிலை இணைப்புக் கிளவி	
	Subordinator	CCS	துணைநிலை இணைப்புக் கிளவி	அனால், ஏனென்றால்
	Quotative		மேற்கோள்	என்று
9	Particles	RP	இடைச்சொல்	உம், ஓ, ஆ
	Default	RP_RPD	வழுநிலை	

	Classifier		பாகுபடுத்தி	
	Interjection	RP_INJ	வியப்பிடைச்சொல்	ஆகா, ஐயோ
	Negation	RP_NEG	எதிர்மறை	
	Intensifier	RP_INTF	மிகப்பான்	மிக
10	Quantifiers	QT	அளவையடை	சிறிது, கொஞ்சம்
	General	QT_QTF	பொது	
	Cardinals	QT_QTC	ஆதார எண்	ஒன்று, இரண்டு
	Ordinals	QT_QTO	முறைமை எண்	ஒன்றாவது, இரண்டாவது, இரண்டாம், மூன்றாம்
11	Residuals	RD	மீதி	
	Foreign word	RD_RDF	அயல் மொழிச்சொல்	புக், மினிட்
	Symbol	RD_SYM	குறியீடு	
	Unknown	RD_UNK	தெரியாதது	
	Punctuation	RD_PUNC	நிறுத்தற்குறி	., ; : ?
	Echowords	RD_ECH	எதிரொலிச்சொல்	(காப்பி) கீப்பீ

5.4.7.6. சர்வேஸ்வரன் மற்றும் மகேசன் என்போரின் விதி அடிப்படையிலான படிநிலை அடையாளக் குழு

சர்வேசரன் மற்றும் மகேசன் என்போர் விதி அடிப்படையிலான படிநிலை அடையாளக் குழு (Hierarchical Tag-set for Rule-based Processing of Tamil Language) முன் மொழிகின்றனர் (Sarveswaran and Mahesan 2014. இவர்கள் பெயர்ச்சொல், வினைச்சொல், முற்றுச்சொல், பகுக்கவியலும் சொல், கூட்டுச்சொல், கடன்சொல், குறியீடு, நிறுத்தற்குறி, பிற, தெரியாதன என பத்து சொல்வகைப்பாட்டு அடையாளக் குழுவை முன்மொழிகின்றனர்.

5.4.8. மூவகைச் சொல்வகைப்பாடு அடையாளப்படுத்தும் வழிமுறைவரைவுகள்

மூன்று சொல்வகைப்பாடு அடையாளப்படுத்தும் வழிமுறைவரைகள் (algorithms) பயன்பாட்டில் உள்ளன (Jurafsky and Martin 2000:289).

1. விதி அடிப்படையிலான சொல்வகைப்பாடு அடையாளப்படுத்தல் (Rule-based tagging)

2. புள்ளியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்தல் (Stochastic tagging)

3.மாற்றம் அடிப்படையிலான சொல்வகைப்பாடு அடையாளப்படுத்தல் (Transformation based tagging)

5.4.8.1.விதி அடிப்படையிலான சொல்வகைப்பாடு அடையாளப்படுதல்

விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் மயக்கம் தருகின்ற சொல் பெயரா வினையா என்று குறிப்பிடுகின்ற கையால் எழுதிய பெரிய தரவுதளங்களை (Database) உள்ளடக்கியது. ஆரம்பக்கட்டத்தில் தானாகவே சொல்வகைப்பாட்டை அடையாளப்படுத்தப் பயன்படுத்தப்பட்ட வழிமுறை வரைவு (algoritham) இரண்டு நிலை அமைப்பால் ஆனது (Jurafsky and Martin 2000:300).

1. அகராதி ஒவ்வொரு சொல்லுக்கும் சாத்தியமான சொல் வகைப்பாட்டைத் தரும்.

2. இரண்டாம் நிலை கையால் எழுதப்பட்ட மயக்கம் நீக்கும் விதிகளை ஒரு சொல்லுக்கு ஒரு சொல் வகைப்பாட்டைத் தருமாறு பயன்படுத்தும்.

ENGTWOL இரண்டு நிலை அமைப்பால் ஆனது. ENGTWOL அகராதி இருநிலை உருபனியல் (Two level Morphology) அடிப்படையில் அமைந்தது. சொல்வகை அடையாளப்படுத்துகை (Parts of Speech Tagging) ஒவ்வொரு சொல்லுக்கும் ஒரு சொல் வகைப்பாட்டை அல்லது பிற சொல் வகை அடையாளத்தைத் தருவதாகும். அடையாளப்படுத்துகை பொதுவான நிறுத்தற்குறிகளுகட்கும் பயன்படுத்தப்படும். கணினிமொழியில் அடையாளப்படுத்துகை (tokenization) போல்தான் இயற்கை மொழியிலும் சொல்லை அடையாளப்படுத்தும் செயல்பாடு அமையும். சொல் வகை அடையாளப்படுத்துகை 'பேச்சு புரிந்து கொள்ளுதல்' (speech recognition), இயற்கை மொழிப் பகுத்தாய்தல் (Natural Language Parsing) மற்றும் தகவல் மீட்டுப்பெறல் (information retrieval) என்பனவற்றில் முக்கியமான பங்கு வகிக்கின்றது. சொல்வகை அடையாளப்படுத்துதலின் உள்ளீடு (input) வழிமுறை வரைவு (Algoritham) சொற்களின் கோர்வைகளும் குறிப்பிட்ட சொல் வகைப்பாட்டு அடையாளக் குழுமம் (Part of Speech tagset) ஆகும். வெளியீடு (output) ஒவ்வொரு சொல்லுக்கும் ஒவ்வொரு நல்ல அடையாளத்தைத் தருவதுதான்.

எ.கா. அந்த <அ.அ> நல்ல <பெஅ> உயர்நிலைப் <பெ> பள்ளியில் <பெ> படிக்கிறான் <ஒன்றன்பால் நிகழ்கால வினைமுற்று>

எளிய எடுத்துக்காட்டுகளுக்குக் கூட ஒவ்வொரு சொல்லுக்கும் அடையாளம் தருவது அவ்வளவு எளிதல்ல.

எ.கா. அவன் <மாபெ> என் <வி, மாபெ> சொல் <வி, பெ> படி <வி, பெ, பிஉ> கேட்டான்

(அஅ – அடைகொளி அடை, மாபெ = மாற்றுப்பெயர், வி = வினை, பெ = பெயர், பி = பின்னுருபு, இ.சொல் = இடைச்சொல்)

POS அடையாளப்படுத்தியின் முக்கியமான வேலை சூழலுக்குத் தக்கவாறு சரியான அடையாளத்தைத் தேர்ந்தெடுத்து மயக்கத்தைத் தீர்ப்பது தான். பெரும்பாலான தமிழ்ச்சொற்கள் எல்லாம் மயக்கம் இல்லாதவை. இருப்பினும் பொதுவாகப் பயன்படுத்தப்படும் சொற்கள் மயக்கம் உள்ளவை.

5.4.8.2.புள்ளியியல் அடிப்படையிலான சொல்வகைப்பாடு அடையாளப்படுத்தல் (Stochastic Part of Speech Tagging)

சாத்தியங்களை அடையாளப்படுத்துவதில் பயன்படுத்துவது மிகப்பழமையானது. பல சாத்திய அடிப்படையிலான அடையாளப்படுத்திகள் வழக்கில் உள்ளன. எ.கா. Hidden Markov Moddel or HMM tagget (Jurafsky and Martin 2000:303). இந்தப் புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தியில் பயன்படுத்தப்படும் உள்ளுணர்வு 'ஒரு சொல்லின் கூடுதலான சாத்தியமான அடையாளத்தை எடு' (Pick up the most likely tag for the word) என்பதாகும்.

தரப்பட்ட வாக்கியத்திற்கு அல்லது சொல் கோர்வைக்கு HMM அடையாளப்படுத்திகள் பின்வரும் வாய்ப்பாடு கூடுதலான சாத்தியத்தியத்தைத் தரும் அடையாளக் கோர்வையைத் தெரிந்தெடுக்கும்.

P(word/ tag)* P (Tag/ previous n tags)

பொதுவாக HMM அடையாளப்படுத்திகள் ஒரு சொல்லைக் காட்டிலும் ஒரு முழு வாக்கியத்திற்கு அடையாளக் கோர்வையைத் தெரிந்தெடுக்கும். நாம் முதலில் எவ்வாறு HMM அடையாளப்படுத்தி ஒரு தனிச்சொல்லுக்கு அடையாளத்தை ஒதுக்குகிறது என்று பார்ப்போம்.

இவ்வகையான ஒரு Bigram-HMM அடையாளப் படுத்தி முந்தைய அடையாளம் t_{i-1} மற்றும் தற்போதைய சொல் w_i என்று தரப்படுகையில் w_i என்ற சொல்லுக்கு மிகச் சாத்தியமான ti என்ற அடையாளத்தைத் தேர்ந்தெடுக்கும்.

$$Ti = \underset{j}{\operatorname{argmax}}\ p\ (tj/t_{i-1},w_i)$$

ஒரு அடையாளப்படுத்தலுக்கு அடிப்படையான HMM சமன்பாடு ti = argmax p(ts/ti –l) p (wi/ti

எ.கா. (Brown Corpus)

Secretariat/ NNP is/ VBZ expected/ XBN to/To race/ VB tomorrow/ NN

People/NNS continue/ VBP to/TO inguire/ VB the/DT reason/NN for/IN the/ DT race/NN for/in Outer/ JJ speace/NN

முதல் எடுத்துக்காட்டில் 'race' ஒரு வினை (VB). 2 வது எடுத்துக்காட்டில் 'race' ஒரு பெயர் (NN) Bigram HMM அடையாளப்படுத்தி அடையாள சிக்கலை அண்மையிலுள்ள சொற்களையும் அடையாளங்களையும் பார்த்துத் தீர்மானிக்கலாம் என்று ஊகம் செய்கிறது.

To/ TO race/???

The/ DT race/???

எவ்வாறு 'race' என்ற நமது எடுத்துக்காட்டு பயன்படுகின்றது என்று பார்ப்போம். நாம் NN-க்கும் VB-க்கும் இடையில் ஒன்றைத் தோ;ந்தெடுக்க வேண்டும். நாம் எதற்குக் கூடுதல் சாத்தியம் இருக்கிறதோ அதைத் தெரிந்தெடுக்க வேண்டும்.

P(VB/To) P (race/VB

P (NN/ To P (race/NN)

பின்வரும் சாத்தியங்களாகும்

P (NN/ To) = :021

P (VB/TO) = .34

5.4.8.3.மாற்றம் அடிப்படையிலான சொல்வகைப்பாடு அடையாளப்படுத்தல் (Transformational based tagging)

மாற்றம் அடிப்படையிலான அடையாளப்படுத்தல் 'Bill tagging' என்று அழைக்கப்படுகிறது. இது பொறிகற்றலுக்கான மாற்றம் அடிப்படையிலான கற்றல் (Transformational- based learning = TBL) அணுகுமுறையாகும் (Jurafsky and Martin 2000:307). இது விதி அடிப்படையிலிருந்தும் புள்ளியல் அடிப்படையில் இருந்தும் விஷயங்களை எடுக்கிறது. விதி அடிப்படையிலான அடையாளப்படுத்திகளைப் போல் TBL-உம் எச்தச் சொற்களுக்கு எந்த அடையாளத்தைத் தரவேண்டும் என்று குறிப்பிடுகின்றது. ஆனால் புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்திகளைப் போல், TBL ஒரு மொழி கற்றல் நுணுக்கமாகும். இதில் தரவிலிருந்து விதிகள் தானாகவே ஊகிக்கப்படுகின்றது. சில HMM Taggers-ஐ போல், TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் நுணுக்கமாகும் (supervised learning technique). இது ஏற்கனவே அடையாளப்படுத்தப்பட்ட பயிற்சி பெற்ற பொருந்தரவை (pre-tagged learning corpus) எதிர்பார்க்கின்றது/ஊகிக்கின்றது. TBL-உக்கு ஒரு குழுமம் அடையாளபடுத்தும் விதிகள் இருக்கின்றது. முதலில் பெருந்தரவு பெரும்பாலும் பயன்படுத்தப்படும் விரிந்த விதிகளை வைத்து அடையாளப்படுத்தப்படும். பின்னர் சில முன்னதான (orignal) அடையாளங்களை மாற்றும் மேலும் திட்ட வட்டமான விதிகள் தெரிந்தெடுக்கப்படும். பின்னர் குறைவான எண்ணிக்கையிலான அடையாளங்களை மாற்றும் மேலும் திட்டவட்டமான விதிகள் பயன்படுத்தப்படும்.

5.4.8.3.1.எவ்வாறு மாற்றம் அடிப்படையிலான கற்றல் விதிகள் பயன்படுத்தப்படுகின்றன

நாம் Brill's tagger பயன்படுத்தும் ஒரு விதியைப்பார்ப்போம். விதிகள் நடைமுறைப் பயன்படுத்துவதற்கு முன் அடையாளப்படுத்தி ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் (mostly likely tag) அடையாளப்படுத்தும். இந்த கூடுதல் உகந்ததான அடையாளங்களை நாம் அடையாளப்படுத்தப்பட்ட பொருந்தரவிலிருந்து பெறுவோம். எ.கா. Brown corpus-இல் race என்பதற்கு கூடுதல் உகந்தது பெயர் என்பதாகும்.

P (NN/race) = .98

P (VB/ race = .02

நாம் முன்னர் பார்த்த 'race' என்பதன் இரு எடுத்துக்காட்டுகளும் NN என்று அடையாளப்படுத்தப்பட்டிருக்கும். முதல் எடுத்துக்காட்டில் NN என்பது தவறான அடையாளம்.

/is/VBZ expected/VBN to/To race/NN tomorrow/NN

2-வது எடுத்துக்காட்டில் 'race' என்பது சரியாக NN என்று அடையாளப்படுத்தப்பட்டிருக்கும்.

The/DT race/NN for/IN outer/JJ space/NN

கூடுதல் உகந்த அடையாளத்தைத் தெரிந்தெடுத்தபின் Brill's tagger அதன் மாற்ற விதிகளை நடைமுறைப்படுத்தும். இது நடக்கையில் 'race' என்பதன் தவறான அடையாளப்படுத்தலில் நடைமுறைப்படுத்தப்படும் விதியைக் கற்றிருக்கும். முன்வரும் Tag To என்றால் NN என்பதை VB என்றும் மாற்றுவதற்கு முன் to/TO என்று வந்தால் இந்த விதி race/NN என்பதை race/VB என்று மாற்றும்.

expected/VBN to/TO race/NN → expected/vbn to/To race/VB

5.4.8.3.2.எவ்வாறு மாற்றம் அடிப்படையிலான கற்றல் விதிகள் கற்கப்படுகின்றன

Brillis TBL வழிமுறை வரைவுக்கு (algorithm) மூன்று முக்கிய நிலைகள் இருக்கின்றன. இது முதலில் ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும். பின்னர் இது ஒவ்வொரு சாத்தியமான மாற்றத்தையும் (possiable transformation) பரிசோதிக்கும். கூடுதல் மேம்பாட்டால் அடையாளப்படுத்தலால் விளைந்த ஒன்றை தேர்ந்தெடுக்கும், இறுதியில் இந்த விதிப்படி தரவை மீள்-அடையாளபடுத்தல் செய்யும். முன்னால் வரும் அடையாளப்படுத்தலை விட குறைந்த மேம்பாடுடைய அடையாளப்படுத்தலான நிறுத்தல் அளவீடு வருவது வரை இந்த மூன்று நிலைகளும் திரும்பித்திரும்பி நிகழும். 2-வது நிலை TBL-க்கு சரியான ஒவ்வொரு சொல்லின் சரியான அடையாளம் தெரியும் என்பதை வேண்டுகிறது. அதாவது TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் வழிமுறைவரைவு (supervised learning algorithm). TBL செயல்பாட்டின் வெளியீடு (outbut) முறைப்படுத்தப்பட்ட மாற்றங்களின்

பட்டியலாகும் (ordered list of transformation) இவை புதிய பெருந்தரவுக்கு (corpus) பயன்படுத்தத்தக்க/நடைமுறைப்படுத்ததக்க அடையாளப்படுத்தல் செயல்முறையை (tagging procedure) உருவாக்கும். கொள்கை அடிப்படையில் சாத்தியமான மாற்றங்களின் குழுமம் எல்லையற்றதாகும். ஆனால் மாற்றங்கள் ஒவ்வொன்றும் வழிமுறை வரைவைக் கடக்கும் போது மிகப்பொருத்தமான ஒன்றை தெரிந்தெடுக்க TBL ஒவ்வொரு மாற்றத்தையும் எடுத்துக்கொள்ள வேண்டும். இவ்வாறு மாற்றங்களின் குழுமத்தை எல்லைப்படுத்த வழிமுறை வரைவுக்கு ஒரு வழி தேவை. இது சிறிய வார்ப்புக்களின் (templates) குழுமத்தை (அருமான மாற்றங்கள்) உருவாக்கிச் செய்யப்படுகின்றது. பின்வருவன பிரில் வார்புருக்களின் (Brills templates) குழுமத்தின் பட்டியலாகும். முன்னர் வருகிற (தொடர்ந்து வருகிற) சொல் ணு ஆக அடையாளப்படுத்தப்படும். இரண்டு சொற்களுக்கு முன்னால் (பின்னால்) வருகிற சொல் Z-ஆக அடையாளப்படுத்தப்படும். முன்னர் வருகிற இரு சொற்களில் ஒன்று Z-ஆக அடையாளப்படுத்தப்படும். முன்னர் வருகிற சொல் Z-ஆகவும் இரு சொற்களுக்கு முன்னால் (பின்னால்) வரும் சொல் W-ஆகவும் அடையாளப்படுத்தப்படும்.

5.5. தொடரியல் பகுப்பான்கள்

முந்தைய பகுத்தாய்வு அணுகுமுறைகள் பெரும்பாலும் அகராதியால் இயக்கப்பட்டன. இதில் அகராதி உருபனியல் தரவையும், பொருண்மையியல் வர்ணனைகளையும், தொடரியல் அறிவையும் கொண்டிருக்கும். பின்னர் இலக்கண வடிவ வாதங்களின் தொடக்கம் பல வேறுபட்ட கொள்கைகள் அடிப்படையிலான பகுத்தாய்வனான்களைத் தந்தது. காஷ்கெட் (Kashket,1986) என்பவர் வார்ல்பிரி என்ற மொழிக்கு சாம்ஸ்கிலின் (Chomsky) ஆளுகை-கட்டுறவுக் கோட்பாடு (government and binding) அடிப்படையில் பகுத்தாய்வானை உருவாக்கினார். இவ்வணுகுமுறையில் தொடர்கள், தேர்வுக் கொள்கைகள், வேற்றுமை அடையாளப்படுத்துகை, வேற்றுமை தருகை, பங்கெடுப்பாளர் தொடர்பு படுத்துகை என்பனவற்றின் வழி உருவாக்கப்பட்டன. பங்கெடுப்பாளர் தகவல் அகராதியில் சேகரிக்கப்பட்டன. சென்குப்தா (Sengupta, et al., 1997) போன்றோர் கப்லன் மற்றும் பிரஸ்னைன் என்போரின் சொல் செயல்பாட்டு இலக்கணத்தை (Lexical Fuhnctional Grammar) வங்காள மொழிக்குத் தகுந்தவாறு மாற்றி ஒரு பகுத்தாய்வானை உருவாக்கினார். சொல் செயல்பாட்டு இலக்கணம் ஆங்கிலம் போன்ற இடம்சார் மொழிகளுக்கு (positional languages) வலுவான வடிவவாதம் ஆகம். இது உறுப்பமைப்பையும் இலக்கண உறவுகளையும் கண்டுகொள்கிறது. செங்குப்தாவின் பகுத்தாய்வான் செய்வினைப்பாடாக உள்ள எளிய வாக்கியங்களுக்கு நல்ல விளவைத் தந்தது. வினிவார்ட்டர் (Winiwarter,1995)ஒன்றிணைப்பு இலக்கணத்தை (unofication grammar)

விரிவாக்கி ஜோமன் மொழிக்குப் பயன்படுத்தினார். சினிவார்ட்டரின் ஒழுங்கமைப்பில் இலக்கண விதிகள் அருவமாக்கலின் பொருத்தமான மட்டத்தில் இலக்கண விதிகள் அதராதிக்கு பங்கெடுப்பாளராகச் சேர்ககப்ப்பட்டன. பெக்கா போன்றோர் (Becker et.al, 1994) V_TAG என்ற கிளை இணைப்பு இலக்கணத்திலிருந்து (TREE adjoining) மாறுபட்ட ஒன்றான பலபெயர் கால பகுத்தாய்வு வழிமுறை வரைவை உருவாக்கினார். இப்பகுத்தாய்வானை கிளை இணைப்பு இலக்கண வடிவவாதம் முற்றிலும் உருவாக்கப்பட்ட மொழிகளுக்கு மட்டுமே பயன்படுத்த இயலும்.

பல பகுப்பான்கள் சார்புக் கொள்கைகள் (dependency principles) அடிப்படையில் உருவாக்ககப்பட்டன. கவிங்டன் (Covington, 2003) உரு சுற்று வரைபடத்தால் இயக்கப்பட்ட சொற்களுக்கு இடையே உள்ள சார்புகளை உருவாக்கி வாக்கியங்களுக்கு அமைப்பைத் தரும் ஒரு பகுப்பானை உருவாக்கினார். இதில் ஒவ்வொரு வில்லும் சொற்களின் சார்புநிலை உறவை உருப்படுத்தம் செய்யும். காரக உறவுகள் வேற்றுமைக் குறிகளுடன் பொருத்தப்படும் பாணினி இலக்கண அடிப்படையில் பாரதி போன்றோர் Gharti, et, al,1993, 2003) கட்டுப்பாடு அடிப்படையிலான பகுப்பான்களின் ஒரு குழுமத்தை உருவாக்கினார். பாரதி ஒழுங்குமுறை பகுக்கும் சிக்கலை முழுஎண் வழியமைப்புச் சிக்கலாக மொழி பெயர்த்துத் தீர்வு காண்கிறது. மூர்த்தி (Murthy,1997) இந்திய மொழிகளுக்கு உரக எச்சத்தொடர் அமைப்பு இலக்கணம் (Universal Clause Structure Grammar) என்ற இலக்கண வடிவ வாதத்தை உருவாக்கினார். இவ்வடிவவாதத்தில் மூன்று மட்ட பகுப்பாய்வு பயன்படுத்தப்படுகிறது. தொடர் அடையாளம் காணல், எச்சத்தொடர் அமைப்பு நிர்ணயம் மற்றும் பங்களிப்பு தருதல். ஒவ்வொரு மட்டத்திற்கும் அதன் பகுப்பாய்வுக்குத் தொடர்புள்ள விதிகள் இருக்கின்றன. உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தில் இத்தொகுதியாக்கம் (modularization) இதைப் பிற மொழிகளுக்கு எளிதாகப் பயன்படுத்த செய்தது. சொல்நிரல் சிக்கலைக் கையாளுவதற்கான ஒரு தீர்மாங்களில் ஒன்று வரிசை மாற்று இலக்கணமாகும் (Permutational Grammar (PG)). (Geg. Ikifssib, et, al., 2001) வரிசை மாற்று இலக்கணம் அடிப்படை தொடரமைப்பு விதிகளை அவற்றின் செயல்பாட்டு உருப்படுத்தங்களுடன் குறிப்பிட்டு பின்னர் ஒரே பொருண்மையைத் தரும் பிற எல்லாத் தொடர்ச்சிகளையும் பெறவேண்டி உறுப்புகளை வரிசைமாற்றம் செய்யும். குமாரசண்முகம் மேலே குறிப்பிட்ட வேறுபட்ட அணுகுமுறைகளின் ஒரு ஒன்றிணைப்பைப் பயன்படுத்துகின்றார். அவர் தொடர்களையும் வாக்கியங்களையும் எச்சத்தொடர்களையும் கண்டுபிடிக்க மூர்த்தியின் உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தைப் போன்று நுட்பங்களைப் (Techniques)

பயனபடுத்துகின்றார். சுதந்திர சொல்நிரல் சிக்கலைக் கையாள வரிசைமாற்று இலக்கணத்தைப் போன்று உறுப்புகளை வரிசை மாற்றுவதில்லை. அதற்குப் பதிலாக வாக்கியங்களின் உறுப்புகளை ஒரு குறிப்பிட்ட நிரலில் வரிசைமாற்றம் செய்கிறார். பாரதியின் ஒழுங்கமைப்பில் (Bharati, 2000) உள்ளது போன்று வாக்கியத்தின் எழுவாயையும் செயப்படுபொருளையும் கண்டுபிடிக்க வேற்றுமை உருபுகளைப் பயன்படுத்துகிறார். தமிழின் தொடரியலை விளக்க பொதுவான தொடரமைப்பு இலக்கணததைப் பயன்படுத்துகிறார்.

5.5.1.தமிழில் தொடரியல் பகுத்தாய்வான்கள்

தமிழ் வாக்கியங்களைப் பகுத்தாய்வதற்கு வேண்டி ஒரு கணினியியல்சார் சட்டகத்தை உருவாக்கி பயன்படுத்துவதாகும். இதற்கு வேண்டி தமிழ்த்தொடரியலை உருப்படுத்தம் செய்ய தொடரமைப்பு விதிகள் உருவாக்கப்பட்டுள்ளன. நாற்பது மிகக் கூடதலாக நிகழும் வினைகளுக்குப் பங்கெடுப்பாளர் அமைப்பு கண்டுபிடிக்கப் பட்டுள்ளது. பகுத்தாயப்பட்ட வாக்கியங்களின் தொடரியல் அமைப்புகள் பகுக்கப்பட்ட கிளைகளில் உருப்படுத்தம் செய்யப்பட்டுளளன. இவ்வடிவில் அவற்றை இயந்திர மொழிபெயர்ப்பு மற்றும் தகவல் பிரித்தெடுப்பு இவற்றை உள்ளடக்கிய பயன்பாடுகளில் நேரடியாகப் பயன்படுத்த இயலும்.

உள்ளீட்டுத் தமிழ் வாக்கியத்தை முக்கியப் பகுப்பான் பகுப்பதற்கு முன் அது பகுப்பாய்வின் பல மட்டங்களைக் கடந்துசெல்ல வேண்டும். கீழே தந்துள்ள படத்தில் செவ்வகங்கள் பகுப்பாய்வு செய்யும் பகுப்பாய்வுத் தொகுதிகளையும் (analyzer modules) வட்டங்கள் அவை பயன்படுத்தும் பொருத்தமான மூலப்பொருள்களையும் உருப்படுத்தம் செய்யும்.

தமிழுக்கு வேண்டி சில பெயர்த்தொடர் பகுப்பான்கள் உருவாக்கப்பட்டுள்ளன. அவை கீழே விளக்கப்பட்டுள்ளன.

5.5.1.1. பாஸ்கரனின் தொடரியல் பகுத்தாய்வான்

முற்றுநிலைத் தானியங்கி மொழியமைப்பு மட்டத்தில் பகுத்தாய்வதற்க்கு முக்கியமான தொழில் நுட்பங்களில் ஒன்றாகும். பாஸ்கரன் (Baskaran, 1984) தமிழ் வாக்கியங்களில் பகுத்துக்குறித்தலுக்கு வேண்டி முற்றுநிலை தானியங்கையைப் பயன்படுத்தினார். இது தமிழ்ப்பல்கலைகழகத்தில் உருவாக்கப்பட்ட ருசியன்-தமிழ் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு வேண்டி இவரால் உருவாக்கபட்டது.

5.5.1.2.எயுகேபிசி ஆய்வுமையத்தின் பெயர்த்தொடர் பகுப்பான்

இவ்வணுகுமுறை விதி அடிப்படையிலானது. இந்நெறிமுறையில் ஒரு தரவுத்தொகுதி எடுக்கப்பட்டு இரண்டோ அதற்கு அதிகமான குழுமங்களாகப் பிரிக்கப்படும். இப்பிரிக்கப்பட்ட குழுமத்தில் ஒன்று பயிற்சித் தரவாகப் பயன்படுத்தப்படும். பயிற்சித் தரவுக் குழுமம் எடுக்கப்பட்டு

பெயர்த்தொடருக்காக மனித முயற்சியால் பகுக்கப்படும். இவ்வாறு ஒரு வாக்கியத்தில் உள்ள பெயர்த் தொடர்களைப் பிரிப்பதற்குப் பயன்படுத்தப்படும் விதிகள் உருவாக்கப்படும். இவ்விதிகள் தொடர் பகுப்பிற்கு அடிப்படையாக அமையும். தொடர் பகுப்பான் வழியமைப்பு இவ்விதிகளைப் பயன்படுத்தி பரிசோதனைத் தரவை பகுக்கும். இவ்விதிகளின் பயன்பாட்டுப் பரப்பு இத்தரவுக்குழுமத்தில் பரிசோதிக்கப்படும். துல்லியமும் மீள அழைத்தலும் இதற்குக் கணிக்கப்படும். விளைவு ஒழுங்குமுறையின் பயன்பாட்டுப் பரப்பை விருத்தி செய்ய கூடதல் விதிகள் தேவையா என்று பரிசோதிக்கப்படும். பின்னர் ஒழுங்குமுறை வேறுபட்ட பிற பயன்பாடுகளுக்குப் பரிசோதிக்கப்படும்.

5.5.1.3.RCILTS-இன் வானவில்

வானவில் இந்திய மொழிகளின் தொழில்நுட்ப முன்னேற்றம் என்ற திட்டத்தின் கீழ் தமிழுக்காகச் செய்யப்பட்ட இயற்கைமொழி ஆய்வுகளில் ஒன்றாகும். இது ஒரு தமிழ் வாக்கியத்தின் தொடரியல் உறுப்புகளை அடையாளம் காண்கிறது. இது பட்டியல் வடிவில் பகுப்புக் கிளைப் படத்தை (parse tree) வெளிப்யிடுகின்றது. இது தனிநிலை வாக்கியங்களையும் கூட்டு வாக்கியங்களையும் கையாளுகின்றது. தனிநிலை வாக்கியங்கள் ஒரு வினையையும் பல பெயர்த் தொடர்களையும் தனிநிலை வினையடைகளையும் பெயரடைகளையும் கொண்டிருக்கும். பல எச்சத் தொடர்கள் உள்ள வாக்கியங்களின் தேர்வுகளில் வானவில் துப்புச்சொற்கள் மற்றும் தொடர்கள் அடிப்படையில் எச்சத் தொடர்களைத் தொடர் குழுக்களாகக் குழுமும். இது தொடரியல் அமைப்பு இலக்கணத்தை உருவாக்கும். வானவில் சொல்நிரலைக் கையாள முன்நோக்குதலைப் பயன்படுத்தும்.

பகுத்தல் என்பது ஒரு வாக்கியத்தை அறிந்துகொண்டு அதற்குத் தொடரியல் அமைப்பைத் தருவதாகும். இது தொடரியல் ஆய்வில் மிகப்பெரிய சிக்கலாகும். இதன் காரணமாக இது இயற்கை மொழி ஆய்வின் ஒரு பெரும்பகுதியாக அமைகிறது.

நிர்ணயிக்கப்பட்ட பகுப்பானின் உள்ளீடு (input) அடையாளப்படுத்தப்பட்ட சொற்களின் நிரலான தொடர்ச்சியும் தொடரமைப்பு விதிகளின் ஒரு குழுமமும் ஆகும். இதன் வெளிப்யீடு (output) வாக்கியத்தின் பகுக்கப்பட்ட அமைப்பாகும். ஒரு வாக்கியத்தின் பகுக்கப்பட்ட அமைப்பு அதன் எழுவாய், செயப்படுபொருள் மற்றும் பயனிலைகளை அறிந்துகொள்ள உதவுகிறது. மேலும், பகுத்துக் குறிக்கப்பட்ட கிளை அமைப்பு பிற தொடர்களையும் அடை செய்யும் தொடர்களையும் அடையாளம் காணப்படுகிறது.

ஒரு மொழிக்கு ஒரு மாதிரிப் பகுப்பான் ஒரு அடையாளப்படுத்தியைப் பயன்படுத்தி வாக்கியங்களின் சொல் வகையை அடையாளப் படுத்துகிறது. பின்னர் பகுத்துக்குறித்தும் வழிமுறை வரைவு (algorithm) வாக்கியத்தைப் பகுத்தாய்ந்து அம்மொழி இலக்கணத்திற்குத் தரப்பட்ட தொடரமைப்பு விதிகளின் அடிப்படையில் பொருத்தமான அமைப்பைத் தரும். பகுத்தாய்வின் போது வழிமுறை வரைவு உறுப்புகளுக்கு இடையேயுள்ள இலக்கண உடன்பாட்டையும் முக்கிய வினைக்கம் செயப்படுபொருள் தொடர்களுக்கும் இடையே உள்ள சார்பு உறவையும் பரிசோதிக்கும்.

5.5.1.4.குமார சண்முகத்தின் தொடரியல் பகுத்தாய்வான்

குமார சண்முகம் தமது எம்.எஸ். பட்டப்படிப்புக்காக 'Parse Representaton of Tamil Syntax' என்ற தலைப்பில் ஒரு ஆய்வேட்டைச் சமர்பித்துள்ளார். இவ்வாய்வேட்டில் இவர் தமிழ்த் தொடரியல் பகுப்பாய்வு பற்றி வரிவான விளக்கம் தந்துள்ளார். சார்பு இலக்கண வடிவவாதம் அடிப்படையில் இவர் தமிழுக்கு ஒரு தொடரியல் பகுத்துக்குறின்பானையும் வடிவமைத்துள்ளார்.

5.5.2. இந்தியமொழிகளுக்கிடையியலன மொழிபெயர்ப்புக்கான தொடரியல் பகுப்பான்

இதில் தொடர்களின் (chunks) வரையறை விளக்கமும் தொடர்களின் வகைகளும் சில தனித்துவம் வாய்ந்த நேர்வுகளும் பேசப்படும். இந்திய மொழிகளுக்கு இடையிலான மொழிபெயர்ப்புத் திட்டத்தில் 11 தொடர்கள் தெரிந்தெடுக்கப்பட்டு அவை தொடர் உறுப்பு அடையாளப் படுத்தலுக்குப் பயன் படுத்தபடுகின்றன.

முக்கியத் தொடர்களின் குழுமம் அவற்றின் உருப்படுத்தமும்

	முக்கிய தொடர் குழுமம்	**உருப்படுத்தம்**
1	பெயர்த் தொடர்	NP
2	வினைமுற்றுத் தொடர்	VGF
3	வினைமுற்றாத் தொடர்	VGNF
4	வினைப்பெயர்த் தொடர்	VGNN
5	வினை எச்சத்தொடர்	VGINF
6	பெயரைடைத் தொடர்	JJP
7	வினையடைத் தொடர்	RBP
8	இணைப்புத் தொடர்	CCP
9	வியப்புத் தொடர்	INTJ
10	எதிர்மறைத் தொடர்	NEGP
11	வெற்றிடத் தொடர்	BLK

தமிழில் தொடரியல் பகுத்துக்குறிப்பான்கள் உருவாக்கும் முயற்சிகள் பல மேற்கொள்ளப்பட்டன. அவற்றில் சில இங்கு பட்டியலிடப்பட்டுள்ளன:

5.6. தமிழில் மொழிநடை ஆய்வி

மொழியின் கட்டமைப்பைக் கருத்தாடல், வாக்கியம், தொடர், சொல் என்ற அலகுகளின் அடிப்படையில் கணிப்பொறி மூலம் அறிய இயலும். மொழி அலகுகள் நடையியல் அடிப்படையிலான சிறப்பு பண்புக்கூறுகளை வெளிப்படுத்தி மொழிநடை வேறுபாட்டை உணர்த்தும். கணிப்பொறியின் உதவியால் வேறுபட்ட நடையில் அமைந்த உரைகளை அல்லது மொழி வழக்குகளை ஆய்ந்து ஒப்பீடு செய்து அவற்றின் நடையியல் சிறப்புப் பண்புகளைத் தெரிந்து கொள்ளலாம். இத்தகைய ஆய்வு படைப்பாளியை இனம் கண்டுக் கொள்ளவும் படைப்பிலக்கியத்தின் காலத்தைக் கணிக்கவும் பயன்படும். கணிப்பொறி உதவியுடன் இத்தகைய ஆய்வை நுணுக்கமாகவும் விரைவாகவும் செய்து முடிக்கலாம். இத்தகைய கணினி வழி தமிழ்மொழி நடை ஆய்வுகள் பல வெற்றிகரமாகச் செய்து முடிக்கப்பட்டுள்ளன. புள்ளியியல் மொழியியல் (Statistical linguistics), அளவீட்டு மொழியியல் (Quantitative linguistics) போன்ற பயன்பாட்டு மொழியியல் பாடங்களின் கோட்பாடுகளும் கொள்கைகளும் மொழியியல் நடை ஆய்வுக்கு உறுதுணைப்புரிகின்றன. மொழி நடையை ஆய உதவும் மென்பொருள்கள் பல கண்டுப்பிடிக்கப்பட்டு பயன்பாட்டில் உள்ளன.

5.7. சுருக்கவுரை

இவ்வியலில் முதலில் வாக்கியங்களையும் தொடர்களையும் பகுப்பாய்வு செய்வதற்கான இலக்கணவடிவாக்கங்கள் பற்றி விரிவாகக் கூறப்பட்டுள்ளது. அண்மையுறுப்பு பகுப்பாய்வு, தொடரமைப்பு இலக்கணம், சூழல் கட்டுப்பாடில்லா இலக்கணம், சூழல் கட்டுண்ட இலக்கணம், மாற்றுநிலை வலைப்பின்னல், பாணினி இலக்கணம், சார்பு இலக்கணம் இவற்றைப் பயன்படுத்தி பகுப்பாய்வு செய்வது பற்றி விளக்கப்பட்டுள்ளது. ஆழமில்லாப் பகுப்பாய்வு குறித்தும் விளக்கம் தரப்பட்டுள்ளது. இரண்டாவதாக வாக்கியங்களைச் சொல் வகைப்பாட்டிற்கு அடையாளப் படுத்துவது குறித்து விளக்கப்பட்டுள்ளது. சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் பற்றிய முந்தைய முயற்சிகள், சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பு, pen சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள், எயுகேபிசி ஆய்வுமையத்தின் தமழுக்கான சொல்வகைப்பாட்டு அடையாளக் குழுமம், IIIT ஹைதாரபாத்தில் உருவாக்கப்பட்ட அடையாளக் குழுமம், AUKBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் மற்றும் அவற்றின் ஒப்பீடு, இந்திய மொழிகளுக்கான BITS சொல் வகைப்பாட்டு அடையாளக் குழுமம் என்பன குறித்து விளக்கப்பட்டது. தமிழுக்கான சொல்வகை அடையாளப் படுத்திகள் பற்றியும் கூறப்பட்டது. மூன்றாவதாக தமிழுக்கான தொடரியல் பகுப்பாய்வு குறித்தும் பகுப்பாய்வான்கள் குறித்தும் விளக்கப்பட்டுள்ளது.

இயல் 6

கணினிப் பொருண்மையியல்

6.1. அறிமுகம்

கணினிப்பெருண்மையில் (Computational semantics) என்பது மொழிக் கூற்றுகளின் பொருண்மையை ஆய்வதாகும். நாம் இங்கு பொருண்மையைக் கண்டுபிடிக்க முறையான உருப்படுத்தங்களையும் கூற்றுக்களிலிருந்து இப்பொருண்மை உருப்படுத்தங்களுக்கு வேண்டி வழிமுறை வரைவுகளிலும் கவனக்குவிப்பு செய்யவேண்டும். நாம் எவ்வாறு கூற்றுக்களின் பொருண்மை தொடர்களின், சொற்களின் மற்றும் உருபன்களின் பொருண்மைகளுடன் தொடர்புடையவை என்று அறிந்திருக்கவேண்டும். பேசுபவர்கள் கேட்பவர்கள் மற்றும் கூற்றுக்கள் காணப்படும் சூழல்கள் இவற்றுடன் தொடர்புடைய சிக்கல்கள் பயன்வழியியலில் தீர்மானிக்கப்படும்.

6.2. பொருண்மையை உருப்படுத்தம் செய்தல்

மொழிக் கூற்றுக்களின் பொருண்மை முறையான அமைப்புகளில் வெளிப்படுத்தப்படும். இதை மொழிகளின் பொருண்மை உருப்படுத்தம் என்பர். இப்பொருண்மை உருப்படுத்தங்கள் ஒலியனியல், உருபனியல் மற்றும் தொடரியல் உருப்படுத்தங்களைப் போன்று பங்களிப்புச் செய்கின்றது.

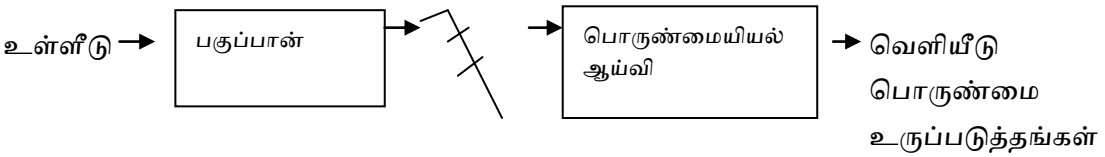
6.3. பொருண்மையியல் ஆய்வு

இங்கு பொருண்மை ஆய்வின் சிக்கல்களுக்கு கணினி அணுகுமுறை கருத்தில் கொள்ளப்படும். முதல் அணுகுமுறை தொடரியலால் இயக்கப்படும் பொருண்மையியல் ஆய்வாகும் (Syntax-Driven Semantic Analysis). இதற்குக் குறைந்த அளவான பயன்பாடுதான் இருக்கின்றது. இது சொற்களஞ்சியத்திலிருந்தும் இலக்கணத்திலிருந்தும் நிலையான அறிவின் அடிப்படையில் உள்ளீடுகளுக்குப் பொருண்மை உருப்படுத்தங்களைத் தரும்.

6.4. தொடரியலால் இயக்கப்படும் பொருண்மையியல்

இது இணைப்பாக்கத்தின் கொள்கை அடிப்படையில் (principle of compostitionality) அமைந்தது. இவ்வணுகுமுறையில் உள்ளுறையும் முக்கிய கருத்து ஒரு வாக்கியத்தின் பொருண்மை அதன் பகுதிகளின் பொருண்மைகளால் இணைக்கப்பட்டது என்பதாகும். வாக்கியத்தின் பொருண்மை முழுவதும் அதில் அடங்கும் சொற்களின் அடிப்படையில் அமையாமல் வாக்கியத்தில் வரும் சொற்களின் நிரல், குழுமுதல் மற்றும் சொற்களுக்கிடையேயுள்ள தொடர்புகள் இவற்றின் அடிப்படையில் அமையும். அதாவது வாக்கியத்தின் பொருண்மை அதன் தொடரியல் அமைப்பின் அடிப்படையில் அமையும் என்பதாகும். எனவே தொடரியலால் இயக்கப்படும் பொருண்மையியல் ஆய்வில் பொருண்மை உருப்படுத்தங்களின் இணைப்பு ஒருவகை இலக்கணத்தால் தரப்படும் தொடரியல் கூறுகளாலும் உறவுகளாலும் வழிக்காட்டப்படும். உள்ளீடு செய்யப்படும் வாக்கியத்தின் தொடரியல் ஆய்வு பொருண்மையியல் ஆய்வின் உள்ளீடாக அமையும். கீழேத் தரப்பட்டுள்ளப்படம் இக்கருதலில்

இருந்து நேரடியாகப் பெறப்படும் பைப்லைன் அடிப்படையிலான அணுகுமுறையை (pipeline-oriented approach) எடுத்துக்காட்டும்:



இவ்வாய்வு பொருண்மை ஆய்வின் உள்ளீடாகத் தரப்பட்டு பொருண்மை உருப்படுத்தத்தை உருவாக்கும்.

6.5. சூழல் கட்டுப்பாடில்லாத விதிகளில் பொருண்மையின் விரிவாக்கங்கள்

சூழல் இக்கட்டுப்பாடில்லாத இலக்கண விதிகள் (Context Free Grammatical rules) பொருண்மையியல் இணைப்புகளால் விரிவுப்படுத்தப்படும். இவ்விணைப்புகளை ஒரு கட்டுமானத்தின் உறுப்புகளின் பொருண்மையிலிருந்து அதன் பொருண்மை உருப்படுத்தம் கணிக்கப்படுகிறது என்பதைக் குறிப்பிடும் கற்றல் விதிகளாகக் கருத இயலும். விரிதாக்கப்பட்ட விதிகளுக்குப் பின்வரும் அமைப்பு இருக்கும்.

6.6. முந்தையப் பகுப்பானில் பொருண்மையியல் ஆய்வை ஒருங்கிணைத்தல்

முந்தைய பத்திகளில் நாம் பொருண்மையியல் ஆய்விகளின் எளிய பைப் லைன் கட்டமைப்பைப் பார்த்தோம். இதில் முழு தொடரியல் பகுப்பாய்வின் கிளைகள் பொருண்மையியல் ஆய்வுக்கு செலுத்தப்படுகிறது. இதன் நோக்கம் என்னவென்றால் இணைப்பாக்க அணுகுமுறை அது தொடர்புப்படுத்தும் தொடரியல் பகுப்பாய்வை வேண்டுகின்றது என்ற கருத்துச்சாயலிலிருந்து வந்ததாகும். இருப்பினும் பொருண்மையியல் ஆய்வை தொடரியல் ஆய்வுக்கு இணையாகச் செய்ய இயலும். முந்தையப் பகுப்பாய்வுடன் பொருண்மை ஆய்வின் இணைப்பு நேரிடையானதாகும். இணைப்பாக்கக் கொள்கை மரபுத் தொடர்களின் (idioms) பயன்பாட்டால் சிக்கலை நேரிடும். மரபுத்தொடர்கள் பிரித்து பொருண்மையை ஆய இயலாத மொழி அலகுகளாகும். இது இணைப்புக் கொள்கைக்கு நேரிடையான விளைவைத்தரும்.

6.7. கணினிச் சொற்பொருண்மையியல்

சொற்களஞ்சியம் (பொருட்புல அகராதி) சொற்களின் பொருண்மை அமைப்பு மற்றும் அவைகளுக்கிடையே உள்ள உறவுகள் எவ்வாறு அமைந்துள்ளன என்பதை விளக்கும். இந்த ஒழுங்குமுறையான பொருண்மையியலில் தொடர்புடைய அமைப்பைப்பற்றிக் விளாக்குவது சொற்பொருண்மையியல் ஆகும். சொற்பொருண்மையியல் சொற்களுக்கிடையேயுள்ள உறவுகளை தொடர் உறவுகளாகவும் அடுக்குறவுகளாகவும் ஆயும். சொற்களுக்கிடையேயுள்ள அடுக்குறவுகள் ஒருபொருள்பன்மொழி, உள்ளடங்கு மொழி, உள்ளடக்கு மொழி, எதிர்மொழி

போன்றவையாகும். மின்சொற்களஞ்சியங்களும் சொல்வலையும் சொற்களுக்கிடையேயுள்ள சொல்லுறவுகள் மற்றும் பொருண்மையுறவுகள் அடிப்படையில் அமைந்தவை.

சொற்பொருண்மையியல் கணினிமொழியியலில் பெரிய ஆய்வுக் களமாக மாறிவருகின்றது. மட்டுமன்றி இது இயற்கைமொழிப் பகுப்பாய்வி மற்றும் ஆக்கிகள் (analyzer and synthesizer) உள்ளடங்கிய பயன்பாட்டின் பல்வேறு வகைகளில் முக்கியப் பங்கு வகிக்கின்றது. சொற்பொருண்மையியல் உளமொழியியலில் இருந்து அறிவு உருப்படுத்தம் (Knowledge Representation (KR)) மற்றும் கணினி அமைப்பு வரைப் பரந்த சிக்கல்களை உட்படுத்தியுள்ளது. இப்பகுதியில் உளமொழியியல், பயன்பாட்டுமொழியல், மொழியியல், கணினிமொழியியல், பயன்பாட்டு வளர்ச்சி போன்றவை சொற்பொருண்மையியலுடன் தொடர்புபடுத்தி விவரிக்கப்பட்டுள்ளன. பின்வரும் கருத்துக்கள் விரிவாகப் பேசப்பட்டுள்ளன:

• உளமொழியியல் (psycholinguistics): மன அகராதிகள், சொல் அலகுகளை அணுகுதல், சொல் அலகுகளின் வடிவு, சொற்களுக்கும் கருத்துருக்களுக்கும் உள்ள தொடர்புகள், அகராதிப்படுத்தல் செயற்பாடு.

• மொழியியல் மற்றும் சொற்பொருண்மையியலின் முறையான நோக்குகள் (linguistics and formal aspects of lexical linguistics): சொற் பொருண்மையியலின் உறவுகள், மூலவகைகள், கருத்துரு உருப்படுத்தங்கள், நிகழ்வு அமைப்புகள், பங்கேற்பாளர் அமைப்பு, சொல்மிகைமை.

• அறிவு உருப்படுத்தங்கள் (knowledge reresentaion): விதிகளின் ஒழுங்குமுறைகள், வகைக்-கீழ்படிவு நடத்துமுறை (type-coercion), மரபுரிமை நோக்கு (aspects of inheritance), மொழியியலுக்கும் உலக அறிவுக்கும் இடையேயுள்ள உறவுகள்.

• பயன்பாடுகள் (applications): பெரிய வடிவிலான அகராதிகளை உருவாக்கிப் பராமரித்தல், பகுத்துக்குறித்தல் மற்றும் உருவாக்கத்தில் அகராதியின் பங்கு, சொல்லறிவு அடிப்படைகள், சொற்தரவின் ஈட்டம்

• நடவடிக்கை நோக்குகள் (operational aspects): மாதிரிகளை வழிமுறைப்படுத்தல் மற்றும் அகராதி ஒழுங்குமுறைகளின் கட்டுகை.

6.7.1. மொழியியல் மற்றும் உளமொழியியல் பார்வையில் சொற்பொருண்மையியல்

உளமொழியியல் மற்றும் மொழியில் இவற்றின் பல கருத்துச்சாயல்கள் சொற்பொருண்மையிலில் பங்குவகிக்கின்றன. இலக்கண வடிவவாதம் நீங்கலாக சொற்பொருண்மையியல் பகுபாய்வுக்கும் உருவாக்கத்திற்கும் மற்றும் வாக்கியம் மற்றும் கருத்தாடல் பொருண்மை உருப்படுத்தம் மற்றும் உருவாக்கம் இவற்றில் கணினிமொழியியலில்

முக்கியப் பங்கு வகிக்கின்றது. கணினிமொழியியலில் சொற்பொருண்மையியலின் முக்கியப் பங்கை சொல் அலகுகள் அவை உருப்படுத்தம் செய்கின்ற சொற்பொருளோடு தொடர்புடைய செய்தியின் பெரும்பான்மையான பாகத்தைக் கொண்டிருக்கின்றது என்ற உண்மையால் விளக்கலாம். இந்தத் தலைப்பின் முதல் பகுதி சொற்பொருண்மையியலில் உளமொழியியல் கருத்துச்சாயல்களின் விருப்பை அறிமுகம் செய்கின்றது. பின்னர் மொழியியல் கருத்துச்சாயல்கள் ஆழமாகப் பேசப்படுகிறது.

6.7.1.1. சொற்பொருள் ஆய்வில் உளமொழியியலாரின் பங்கு

உளமொழியியல் ஆய்வு எவ்வாறு கருத்துருக்கள் நினைவகத்தில் ஒழுங்கு படுத்தப்பட்டுள்ளன மற்றும் மன அகராதியில் எவ்வாறு இந்தச் செய்திகள் தொடர்புகொள்ளப் படுகின்றன என்பன குறித்து நல்ல கருத்துக்களை நமக்குத் தருகின்றன. மேலும் புரிந்துகொள்வதிலும் உருவாக்கத்திலும் மன அகராதியிலிருந்து மீட்கப்படும் உகந்த செய்திகளின் ஆய்வு கணினி அகராதிகளில் காணப்படும் செய்தியைப் பற்றித் துப்புக்களைத் தருகின்றது. உளமொழியியலின் கருத்துக்களின் முக்கியத்துவத்தைச் செயற்கை அறிவில் (artificial intelligence) (குறிப்பாக அறிவு உருப்படுத்தத்தில்) அதன் தாக்கத்தால் எடுத்துக்காட்டலாம். உளமொழியியல் அறிவு உருப்படுத்தத்தின் (knowledge representation) மாதிரிகளின் திட்டத்தில் முன்னணியான பங்கு ஏற்கின்றது. உளமொழியியல் புலனறிவு உளவியலில் (cognitive psychology) (குறிப்பாக புலனறிவு அறிவியிலில்; (cognitive science)) பெருமளவு ஒன்று சேர்ந்துள்ளது.

ஒரு மன அகராதி நம் மூளையில் சேசரித்துவைத்துள்ள சொல்லன்களின் பட்டியலிலிருந்து அவற்றை எடுத்துத் தரவேண்டும். இதனால் பின்வருவன சாத்தியமாகும்:

- நபர், பொருட்கள், இடங்கள், நிகழ்ச்சிகள், மற்றும் உணர்ச்சிகள் என்பனவற்றைப் பெயரிட்டு விளக்குதல்
- நம் எண்ணங்களை சொல்லாக்கம் மற்றும் இலக்கவாக்கம் வழி வெளிப்படுத்தல்
- நாம் கேட்கும் படிக்கும் அல்லது உச்சரிக்கும் சொற்களைப் புரிதல் அல்லது உருவாக்குதல்

நாம் முதலில் மன அமைப்பின் மாதிரியைத் தந்து மன அணுகலின் வேறுபட்ட அணுகுமுறைகளைக் காண்போம்.

6.7.1.1.1. காலின் மற்றும் குல்லியன் மாதிரி (Models of Collins and Quillian)

காலின் குல்லியன் மாதிரியைப் பின்வரும் பகுதிகளை உள்ளடக்கிய ஒரு வலைபோன்ற அமைப்பில் தரலாம் (Saint-Dizoer and Viegas 1995: 3):

• வகைக் கணுக்கள் (type nodes) (தாவரம்)

• வருகைக் கணுக்கள்; (occurrence nodes) (அமைப்பு, உயிருள்ளவை)

• கணுக்களை ஒன்றுடன் ஒன்றை இணைக்கும் தொடர்பு உறவுகள் (மற்றும், அல்லது)

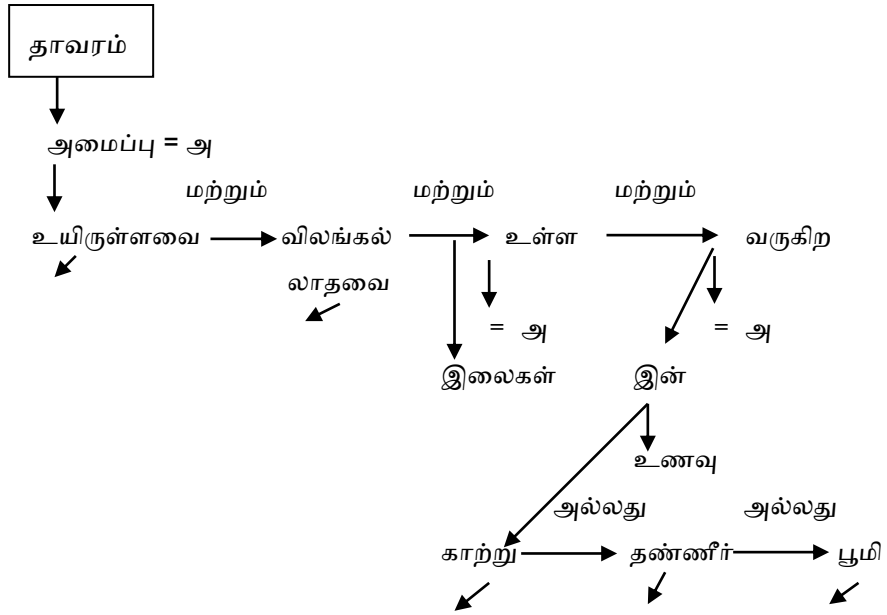
அடிப்படையில் சொல் வலையானது கருத்துருக்களை உட்படுத்தம் செய்யும் கணுக்கள் மற்றும் கணுக்களுக்கிடையேயுள்ள ஈரிணை உறவுகளான விளிம்புகள் உள்ள வரைபடங்களின் (graphs) கருத்துச் சாயலால் ஆனது. இந்த மாதிரியில் சொற்கள் மூலமானதல்ல. பின்வரும் படத்தில் தாவரம் என்பதன் பொருள் உருப்படுத்தம் செய்யப்பட்டுள்ளது: பெரும்பாலும் இலைகள் உள்ள காற்று, தண்ணீர், பூமி இவற்றிலிருந்து உணவைப் பெறுகின்ற விலங்கல்லாத உயிர் அமைப்பு (Saint-Dizoer and Viegas 1995:3).

இம்மாதிரியான மாதிரி களஞ்சிய அறிவில் கவனம் செலுத்துகின்றது. பிற மாதிரிகள் பின்வருவனவற்றை வேறுபடுத்துகின்றன:

1. நீண்டகால நினைவகம் (long term memory)

2. தற்போதைய நினைவகம் (வேலைக்கான நினைவகம் அல்லது குறுகியகால நினைவகம்) (Current memory/working memory or short term memory)

(1)

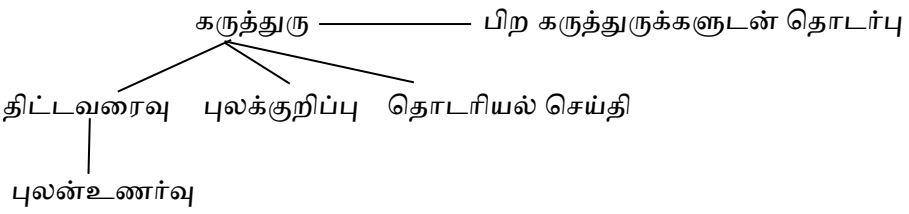


காலின் குல்லியன் மாதிரி சிக்கனத்தைக் கடைபிடிப்பது. நினைவக ஒழுங்குத் திறமையைப் பொறுத்து அமைகின்றதல்லாமல் சிக்கனத்தைப் பொறுத்து அமையாது என்ற நிலையில் காலின் குல்லியன் மாதிரி குறையுள்ளது.

6.7.1.1.2. மில்லர் மற்றும் ஜாண்சன்-லைர்ட் மாதிரி (The model of Miller and Johnson-Laird)

இவர்கள் புலனுணர்வுக்கும் (perception) மொழிக்கும் உள்ள உறவில் விருப்பம் காட்டினார்கள். பொருள்களையும் நிகழ்வுகளையும் சொல்-புலக்குறிகளுடன் (lexical labels) இணைப்பதை அனுமதிக்கும் செயற்பாங்கைக் காண முயற்சித்தார்கள். அவர்களைப் பொறுத்தமட்டில் புலனுணர்வு என்பது அமைப்புமுறையில் அமைந்த செயல்பாடு. அதன் விளைவு ஒரு பொருளிலிருந்து வேறுபடுத்தும் குணங்களைப் (வடிவம், நிறம், பரிமாணம், இருப்பு) பிரித்தெடுக்க உதவும். புலனுணர்வு திட்டவரைவுகளைக் (schemas) குறிப்பிட்டதால் கருத்துருவாக்கச் செயல்பாடகக் கருதப்பட்டது. (திட்டவரைவு என்பது ஒரு கருத்துருவுடன் சேர்க்கப்பட்ட தேவையான புலன் உணர்வுச் செய்திகள் ஆகும்.) (Saint-Dizoer and Viegas 1995:3).

(2)



கருத்துரு மட்டம் திட்டவரைவுகளுக்கும் சொற்களுக்கும் இடையே இணைப்பானாகச் செயல்படுகின்றது. ஒருசொல் கருத்துரு பின்வருவனற்றைக் கொண்டிருக்கும் (Saint-Dizoer and Viegas 1995:4):

1. புலக்குறிப்பு (label)
2. திட்டவரைவு (Schema)
3. புலக்குறிப்பின் தொடரியல் நடத்தையை ஆளும் விதிகள் (rules governing the syntactic behaviour of the label)

திட்டவரைவு செயல்பாடு மற்றும் புலன்உணர்வுச் செய்திகளை இணைக்கின்றது. திட்டவரைவு நேரடியான புலனுணர்வு தாக்கமில்லாத நினைவகச் செய்திகளையும் உள்ளடக்கலாம். சொற்களின் மொழியியல் பொருள் அவைகளுக்கிடையிலான வேறுபட்ட உறவுகளால் ஆக்கப்படுகின்றது. ஒரு மனஅகராதி இரண்டு கூறுகள் உடையதாக இருக்கும்:

1. வகுப்பு உள்ளடங்கல், முழுமை-பகுதி உறவுகள், உறுப்பமைவு மற்றும் அடக்குறவுகள் ஆகிய சொல்லன்களுக்கிடையே உள்ள உறவுகள் அடங்கிய சொல் கூறு (**lexical component**).

2. பல புலனுணர்வுக் கூறுகள், குறைந்த தொடரியல் செய்தி, நினைவுசார் செய்தி (உணர்ச்சி, செயல்பாடு) அடங்கியன அடங்கிய புலனுணர்வுக் கூறு (**perceptual component)**).

சமீபகால ஆய்வுகள் நீண்டகால நினைவுக்கும் குறைந்தகால நினைவுக்கும் வேறுபாடு இல்லை என்பதைக் காட்டுகின்றன.

6.7.1.1.3. புரிதலில் சொல்சார் அணுகல் (Lexical access in recognition)

புரிதலில் சொல் அணுகல் சிக்கலைப் பின்வருமாறு சுருங்கக் கூறலாம் (Saint-Dizoer and Viegas 1995:4):

1. எவ்வாறு நாம் தனிச்சொல்லாகவோ கூட்டுச்சொல்லாகவோ வரும் சொல்லைப் படிக்கவோ கேட்கவோ செய்யலாம்?

2. எவ்வாறு நாம் அவற்றை மனச் சொற்களஞ்சியத்திலிருந்து மீட்கலாம்?

நாம் பல்சொல் கூற்றுக்களைப் புரிந்துகொள்கையில் தனிச்சொற்களை அடைவது மட்டுமன்றி சொல் பொருண்மையியல் மற்றும் வாக்கியப் பொருண்மை அடிப்படையில் அவற்றிற்கு இடையே உள்ள உறவுகளையும் கணக்கில் எடுத்துக்கொள்ள வேண்டும்.

6.7.1.1.4. உருவாக்கலில் சொல்சார் அணுகல் (Lexical access in production)

உருவாக்கலில் சொல்சார் அணுகல் சிக்கலைப் பின்வருமாறு சுருங்கக் கூறலாம் (Saint-Dizoer and Viegas 1995:4):

1. எவ்வாறு நாம் தரப்பட்ட ஒரு பொருளையோ நிகழ்வையோ சரியான சொல் புலக்குறிப்புடன் தொடர்பு படுத்தலாம்.

2. எவ்வாறு கூற்று உருவாக்க வேலைச் சூழலுக்குள் சொல் மீட்பு நடக்கிறது?

திட்ட நிலையில் இரண்டு மட்டங்கள்:

1. செயல்பாட்டு மட்டம் (functional level): இங்கு சொற்கள் இலக்கண மற்றும் பொருண்மைச் செய்திகளைக் கொண்டிருக்கும் அருவ வடிவில் காணப்பெறும்.

2. இருப்பு மட்டம் (position level): இங்கு இலக்கண உருபன்கள் மற்றும் ஒலியனியல் வடிவுகள் சொருகப்படுகின்றது மற்றும் உருவாக்கத்தின் நிரல் தரப்படுகன்றது.

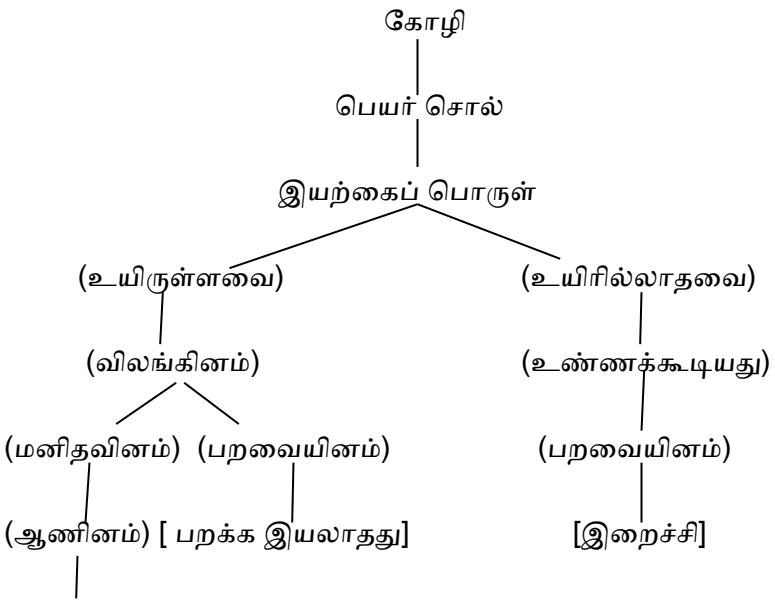
6.7.1.1.5. தானியக்க அல்லது கட்டுப்பட்ட செயற்பாங்கா? இணையான அல்லது தொடரான நடைமுறையா? (Automatic or controlled process? Parallel or serial treatment)?

உளமொழியியலார் நாம் சொல் அணுகலில் கட்டுண்ட செயற்பாங்கைக் கைக்கொள்ள வேண்டுமா தானியக்க செயற்பாங்கைக் கைக்கொள்ள வேண்டுமா என்ற கேள்விக்கு இறுதியான விடை அளிக்கவில்லை. மன அகராதி தன்வயமானது. மிகுந்த வளைவுசுழிவுகளை அனுமதிக்கிற பல்லணுகல் வழிகள் மற்றும் நல்ல செயல்பாட்டின் சீர்மையுடன் சிக்கன அடிப்படையில் அல்லாமல் திறமைக் கொள்கையால் கையாளப்படுகின்றது. அகராதிகள் அமைப்பாக்கம் செய்ப்பட்டும் ஒழுங்கு செய்யப்பட்டும் இருக்கின்ற வழிமுறைகள் அவைகள் எவ்வாறு அணுகப்படுகின்றன என்பதில் பாதிப்பை ஏற்படுத்துகின்றன.

6.7.1.1.6. கூறாய்வுப் பொருண்மையில் (Componential Semantics)

கூறாய்வில் ஒரு சொல்லின் புலன்/பொருள் அம்மொழியலுள்ள பிற சொற்களிடமிருந்து அதை வேறுபடுத்தும் ஒரு குழுமப் பண்புக்கூறுகள் என்று வரையறை செய்யப்படுகின்றது. இது ஜெம்ஸ்லெவ் என்பவரால் தொடங்கப்பட்டு காட்ஸ் மற்றும் பாடர் என்பர்களால் முன்னேற்றப்பட்டது. கீழ்வரும் எடுத்துக்காட்டில் வாத்து, தொடரியல் அடையாளங்கள், (பொருண்மையியல்; அடையாளங்கள்), [பொருண்மையியல் வேறுபடுத்திகள்], <தொடரியல் அல்லது பொருண்மையில் வரையறை> போன்ற கருத்துச்சாயல்களால் ஆயப்படுகின்றது (Saint-Dizoer and Viegas 1995:8).

(3)



வேறுசில ஆய்வுகள் மூலவகைக் கருத்துச்சாயலை அடிப்படையாகக் கொண்டது. பின்வரும் அட்டவணைகள் கோபுரம் மற்றும் வீணை என்பவைகளுக்கான மூலவகை அடைகளைக் காட்டுகின்றது (Saint-Dizoer and Viegas 1995:8).

(4)

கோபுரம்	
உயரம்	
காலம்	
அழகு	
வலிமை	
விலை	
இலாபம்	
….	
எடை	
….	
இரசாயனப்பொருள்கள்	

(5)

வீணை	
ஒலி	
வடிவு	
நிறம்	
அழகு	
விலை	
செய்பொருள்	
...	
எடை	
...	

6.7.1.1.7. செயல்முறைப் பொருண்மையில் (Procedural semantics)

செயல்முறைப் பொருண்மையில் மில்லர் (Miller) மற்றும் ஜாண்சன் லைர்ட் (Johnson Laird (1976) என்போரால் உருவாக்கப்;பட்டது. இதில் ஒரு சொல்லின் பொருள் அதன் பயன்பாட்டில் அடங்கும் செயல்முறைகளால் வரையறை விளக்கம் செய்யப்படுகின்றது. சொல் ஒரு நிலையான இருப்புப்பொருட்களாகக் (entities) (ஒரு குழுமப்ப பண்புக்கூறுகள் அல்லது தேற்றக்கூற்றுகள்) கருதப்பெறாமல் துணை-நடைமுறை ஒழுங்குமுறையாக் (sub-routine) கருதப் பெறுகின்றது. ஒரு வாக்கியத்தின் புலப்பொருள் கணினி வழியமைப்புக்கு இணையானதாகக் கருதப்படும். மில்லரும் ஜாண்சன் லைர்டும் did Lucy bring the desert? ? என்ற கேள்வியை ஆய ஒரு குழும கட்டளைகள் தருகின்றார்கள் (Saint-Dizoer and Viegas 1995:9)

(6)

F(x,y) என வரும் ஒரு நிகழ்ச்சியைக் கண்டுபிடிக்கவும்.

F என்பதற்கு << F(x,y) >> என்ற மதிப்பைத் தரவும்,

x என்பதற்கு << bring >> என்ற மதிப்பைத் தரவும்,

y என்பதற்கு << desert >> என்ற மதிப்பைத் தரவும்.

விவரணை கிடைத்தால், விடை << yes >> என்பதாகும், ஒன்றும் கிடைக்கவிலை என்றால், விடை << I don't know >> விவரணையை மறுக்கும் செய்தி கிடைத்தால், விடை << no >> .

6.7.1.2. மொழியலில் சொற்பொருண்மையில் (Lexical semantics in Linguistics)

பின்வருவன மொழியிலில் சொற்பொருண்மையிலின் முக்கியமான பரப்புகளாகும் (Saint-Dizoer and Viegas 1995:9-25).

1. பங்கெடுப்பாளர் அமைப்பு (Argument structure)
2. மையப்பொருள் பாத்திரங்கள் (Thematic roles)
3. தேர்வுக்கட்டுப்பாடுகள் (Selectional restrictions)
4. சொற்பொருண்மையியலின் முக்கிய உறவுகள் (Lexical semantics essential relations)
5. மூலப்பொருள் ஆய்வுகள் (Ontologies)
6. சொல் கருத்துரு உருப்படுத்தங்கள் (Lexical semantics essential relations)
7. நிகழ்வு அமைப்பு (event structure)
8. தகுதி அமைப்பு (qualia structure)

6.7.1.2.1. பங்கேற்பாளர் அமைப்பு (Argument structure)

ஒவ்வொரு பயனிலையும் அது தேவைப்படும் பங்கெடுப்பாளர்களைக் குறிப்பிடும் ஒரு பங்கேற்பாளர் அமைப்பைக் கொண்டிருக்கும் (Grimshaw, 1990). பங்கேற்பாளர்கள் ஒரு பயனிலையால் விளக்கப்படும் செயலிலோ நிலையிலோ கட்டாயமாக உட்படுத்தப்படவேண்டிய அலகுகளை உருப்படுத்தம் செய்யும். அடைகள் (Adjuncts) செயல் அல்லது நிலை விளக்கங்களில் கூறத்தக்க விதத்தில் பங்கு கொண்டாலும் பங்கேற்பாளர் விளக்க அமைப்பில் உட்படுத்தப்படுவதில்லை. வினைகளுக்கும் சில பெயர்களுக்கும் பெயரடைகளுக்கும் முன்னுருபுகளுக்கும்/பின்னுருபுகளுக்கும் பங்கேற்பாளர் அமைப்பு உண்டு. பங்கேற்பாளர் அமைப்பு உள்வகைப்பாட்டுச் சட்டகம் மற்றும் மையப்பொருள் பங்களிப்புகள் (thematic roles) என்ற கருத்துச்சாயல்களுடன் தொடர்புபடுத்தப்படுகின்றது என்றாலும் பங்கேற்பாளர் அமைப்பு பங்கேற்பாளர்களின் தொடரியல் வகைப்பாட்டைப்பற்றி ஒன்றும் கூறுவதில்லை. ஒரு வினையின் பங்கேற்பாளர் அமைப்பு அவ்வினைவரும் தொடரியல் கட்டுமானத்தில் எத்தனை பங்கேற்பாளர்கள் இருக்கவேண்டும் என்பதை விளக்கும்.

(7)

கொடு: 3 (அல்லது, மாறாக: கொடு: X,Y,Z)

நினை: 2

வாதிடு: 2

தூங்கு: 1

வினைகளுடன் தொடர்புடைய அடைகளான காலம், இடம் அல்லது முறைமை போன்றவை கூடுதல் செய்திகளைப் பயனிலைக்குத் தரும். ஆனால் அவைகள் பங்கேற்பாளர் அமைப்பில் வருவதில்லை. சில சமயங்களில் பங்கேற்பாளர் விடப்படலாம்.

(9) ராமன் புத்தகம் கொடுத்தான்.

வினைகளைப் போல் பல பெயர்களுக்கும் அடைகளுக்கும் பின்உருபுகளுக்கும் பங்கேற்பாளர் விளக்க அமைப்பு உண்டு. அவற்றின் பங்கேற்பாளர்கள் விருப்பாக வருவதோடு வாக்கியத்தில் உருப்படுத்தம் செய்யப்படாதும் வரலாம். எடுத்துக்காட்டு:

(10) அழிவு: பெயர், 1 (நகரத்தின் அழிவு)

முயற்சி: பெயர், 2, (போவதற்கு X -இன் முயற்சி)

வரவு: பெயர், 1 (X -இன் வரவு)

திறமையான: பெயரடை, 1 (திறமையான X)

வெளியே: பின்உருபு, 2 (X Y -க்கு வெளியே)

இடையில்: பின்உருபு, 3 (X Y-க்கும் ணு-க்கும் இடையில்)

6.7.1.2.2. மையப்பொருள் பங்களிப்புகள் (Thematic roles)

பங்கேற்பாளர்களுக்கும் பயனியலைக்கும் உள்ள உறவை மையப்பொருள் பங்களிப்புகளால் உருப்படுத்தம் செய்யலாம். எடுத்துக்காட்டாக மூன்று ஏற்புகள் பெறும் *கொடு* என்பதன் முதல் பங்கேற்பாளர் செய்பவர் என்ற பங்களிப்பையும் இரண்டாவது பங்கேற்பாளர் செயலுக்குள்ளாகுபவர் என்ற பங்களிப்பையும் பெறும். மையப்பொருள் பங்களிப்புகளின் வரையறை விளக்கம் பெரும்பாலும் முழுமையாக அமைவதில்லை. மிகமுக்கியமான மையப்பொருள் பங்களிப்புகளின் வரையறை விளக்கங்கள் கீழே தரப்பட்டுள்ளன.

1.செய்பவர் (agent): செயலைச் செய்வதாகவோ செயல்படுத்துவதாகவோ பயனிலையால் நிர்ணயிக்கப்படுகின்ற பங்கெடுப்பாளர். எ.கா.

(12) உண், இடி, கவனி, கொடு என்பதன் முதனிலை பங்கேற்பாளர்.

2.செயலுக்குள்ளாபவர் (patient): செயலுக்கு உள்ளாகும் அல்லது செயலால் பாதிக்கப்படும் பங்கேற்பாளர்.எ.கா.

(13) கொல், உண் என்பதன் இரண்டாம்நிலைப் பங்கேற்பாளர்.

3. கருப்பொருள் (theme): இடம், நிலைமை அல்லது நிலை மாற்றுகின்ற அல்லது ஒரு நிலையில் அல்லது இருப்பில் இருக்கின்ற பங்கெடுப்பாளர்.எ.கா.

(14) கொடு என்பதன் முதல்நிலை பங்கேற்பாளர், நட, இற என்பதன் பங்கேற்பாளர்.

4. அனுபவிப்பவர் (experiencer): ஒன்றைப்பற்றி தெரிவிக்கப்படும் அல்லது பயனிலையால் வெளிப்படுத்தப்பட்ட ஒரு உளநிலையை அனுபவிக்கும் பங்கெடுப்பாளர். எ.கா.

(15) காதலி என்பதன் முதனிலைப் பங்கேற்பாளர், வெறுபூட்டு என்பதன் இரண்டாம் நிலைப் பங்கேற்பாளர்.

5. மூல இடம் (source) : நகர்வு எங்கிருந்து நிகழுகிறதோ அது. எ.கா.

(16) நீங்கு என்பதன் இரண்டாம் நிலைப் பங்கேற்பாளர்.

6. இலக்கு இடம் (Goal)(: நகர்வு எதைநோக்கி நிகழ்கிறதோ அது. எ.கா.

(17) வந்துசேர் என்பதன் இரண்டாம் நிலைப் பங்கேற்பாளர்.

7.இடம் (Location) : பயனிலையால் விளக்கப்படும் செயலோ நிலையோ நிகழுமிடம். எ.கா.

(18)விழு என்பதன் இரண்டாம் நிலைப் பங்கேற்பாளர்.

8. பயன்பெறுபவர் (Benefactory): வினையால் வெளிப்படுத்தப்படும் செயலால் பயன்பெறும் பங்கேற்பாளர். எ.கா.

(19) கொடு என்பதன் இரண்டாம்நிலைப் பங்கேற்பாளர். மையப்பொருள் பாத்திரங்களின் படிநிலை ஒழுங்கு (செய்பவர் (அனுபவிப்பவர் (குறியிடம்/மூலவிடம்/இடம் (மையப்பொருள்/செயலுக்குள்ளாபவர்)

(20) கொடு (செய்பவர் (குறியிடம்(கருப்பொருள்)))

(21) சந்தி: (செய்பவர் (கருப்பொருள்))

(22) பயப்படு: (அனுபவிப்பவர் (மையப்பொருள்)

மையப்பொருள் மூலப்பங்களிப்புகளை நோக்கி (Toward thematic proto-role)

குரூஸ் (Cruse, 1986) செயலி என்பதற்குப் பின்வரும் சிதைத்தலைச் செய்கின்றார்:

(23) செய்பவர்: +விருப்பச்செயல், +செயன்மை, +செயலூக்கம் +முதன்முயற்சி

டவுத்தி (Dowty , 1991) மையப்பொருள் கருத்துச்சாயலைப் பின்வருமாறு மூன்று வகைகளாகக் குறிப்பிடுகின்றார்:

1. கூடுதல் மையப்பொருள் (increment theme)

எ.கா. கட்டுரை எழுது

2. முழு மையப்பொருள் (holistic theme)

எ.கா. சென்னைக்குப் போ.

3. மூல இட மையப்பொருள் (source theme)

எ.கா. கோப்பைக் குறுந்தகட்டில் நக்லெடு.

டவுத்தி செய்பவரையும் செயலுக்குள்ளாகுபவரையும் மூலப் பாத்திரங்களாகக் கொண்டு அவற்றைப் பல சிறப்புப் பண்புகளால் விவரணை செய்கிறார்:

(24) செய்பவர்:

+விருப்பச்செயல் (ராமன் வேலை செய்கிறான்.)

+புலனணர்வு (ராமனுக்கு அது தெரியும்.)

+ஒரு நிகழ்வை நிகழச்செய்தல் (தொலைபேசி மணி தடை செய்தது.)

+சலனம் (காற்று கூரை பிய்ய வீசியது.)

குறிப்பிடப்படும் (referent) பொருள் பயனிலையால் விளக்கப்படும் செயலிலிருந்து சுதந்திரமாக இருக்கும்.

செலுக்கு ஆளாகுபவர்:

+நிலைமாற்றம் (ராமன் கதவைத் திறந்தான்)

+கூடுதல் மையப்பொருள் (ராமன் புத்தகம் எழுதுகிறான்)

+காரணத்தாக்க நிகழ்வு (புகைபிடிப்பது கான்சரை உருவாக்கும்.)

+நிலையானது, மூல-செய்பவரை ஒப்பிடுகையில் (பேருந்து நிறுத்துமிடத்துக்குள் புகுந்தது.

6.7.1.2.3. தேர்வுக் கட்டுப்பாடுகள் (Selectional restrictons)

தேர்வுக்கட்டுப்பாடுகள் பங்கேற்பாளர்களின் பொருண்மைக் கூறில் கூடுதலான பயன்வழியியல் அளவையை அறிமுகப்படுத்தும். தோ;வுக்கட்டுப்பாடுகள் இயற்கை மொழியியல் ஆய்வில் பெரிதும் பயன்படுத்தப்படுகின்றன. இவை செற்கை அறிதிறன் அறிவு உருப்படுத்த நுண்மையை அடிப்படையாக்கொண்டது. ஒரு பயனிலையின் ஒவ்வொரு பங்கெடுப்பாளரையும் ஒரு பட்டியல் பொருண்மைப் பண்புக்கூறுகளோடு தொடர்புபடுத்துவது இதன் அடிப்படைக் கொள்கையாகும் (Saint-Dizoer and Viegas 1995:14).

(14)

• [Fi] ஒரு தனிப்பட்ட பொருண்மைக் கூறு.

எ.கா. மனிதவினம், விலங்கினம், உண்ணக்கூடியது போன்றவை.

• [F1^F2^...Fn] கட்டாயம் வரவேண்டிய கட்டுப்பாடுகளின் இணைப்பை வெளிப்படுத்தும் அலகுகளின் இணைப்பு.

• [F1vF2...vFn] கட்டுப்பாடுகளை வேறுபடுத்தல். ஏதாவது ஒன்று கட்டாயம் வரவேண்டும்.

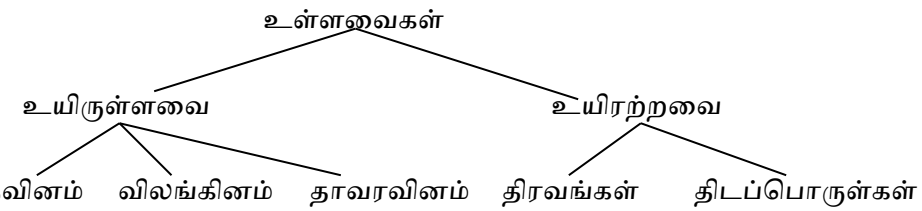
• இணைப்புகள் மற்றும் வேறுபடுத்தல் என்பதன் கலவை.

எடுத்துக்காட்டாக குடி என்ற வினை பின்வரும் பொருண்மைப் பண்புக்கூறுகளைக் கொண்ட செயப்படுபொருளை வேண்டும்.

(15) [திரவம்^உண்ணக்கூடியது

தேர்வுக்கட்டுப்பாட்டில் பயன்படுத்தப்படும் பொருண்மைவகைளை வகை/துணைவகை உறவால் அமைப்பாக்கம் செய்யலாம் (Saint-Dizoer and Viegas 1995:14).

(16)



6.7.1.2.4. சொற்பொருண்மை உறவுகள்

இங்கு நாம் சொல் அமைப்பாக்கத்தின் அதிக உலகளாவிய நிலையையும் சொற்களை அமைப்புசெய்யும் வேறுபட்ட மிக அறியப்பட்ட உறவுகளையும் கருத்தில் கொள்வோம். சொற்பொருண்மை உறவுகள் சொற்பொருண்மையில் முக்கிய பங்கு வகிக்கின்றது மற்றும் சொற்பொருண்மையியல் இயற்கைமொழிப் புரிந்து கொள்கையிலும் உருவாக்கத்திலும் பல நிலைகளில் ஊடாட்டம் செய்கின்றது. சொற்பொருண்மையியல் அறிவு அடிப்படைகளின் (lexical semantics knowledge bases) அமைப்பாக்கத்தில் மையக்கூறுகளாகும். கீழே தரப்பட்டுள்ளது குரூசிடமிருந்து (Cruse, 1986) எடுத்தாளப்பட்டுள்ளது.

அடிப்படை உறவுகள் (congruence relations)

E1, E2 என்ற இருப்பான்களை குறிப்பிடும் W1, W2 என்ற சொற்கள் பின்வரும் உறவுகளில் வரும்:

ஒத்தல் (Identity): E1 = E2

உள்ளடங்கல் (Iclusion): E2 என்பது E1 என்பதில் உட்படுத்தப்பட்டிருக்கும்

மேலுறல் (Overlap): E1, E2 என்பன ஒரு வெற்றிடமற்ற மேலுறலைக் கொண்டிருக்கும்; ஆனால் ஒன்று மற்றொன்றில் உட்படுத்தப்பட்டிருக்காது.

வேறுபடல் (Disjunction): E1, E2 என்பன பொதுவான கூறுகளைக் கொண்டிருக்காது.

இந்த உறவுகள் வகை/துணைவகை உறவு போன்ற வேறுபட்ட வகையில்படும் சொல்சார் கட்டமைப்பைக் கொண்டிருக்கும்.

படிநிலை உறவுகள் (Hierarchical relations)

மூன்று முக்கியமான படிநிலை உறவுகள் உள்ளன: வகைபாடுகள் (Taxonomies), சினைமுமைகள் (meronomies) விகித வரிசைகள்.

வகைப்பாடுகள் (Taxonomies),

வகைப்பாடு என்பது "ஆக இரு" (isa) என்ற மிக அறியப்பட்ட உறவாகும்; இது ஒரு குறிப்பிட்ட வகையை (உள்ளடங்குமொழியம்/hyponym) ஒரு பொதுவகையுடன் தொடர்புபடுத்தும். வகைப்பாடு பின்வரும் மொழியியல் சோதனைகளால் பண்பாக்கம் செய்ய இயலும் வகை/துணைவகை உறவுகளை உட்படுத்தும்.

பின்வரும் கூற்றுகள் சரி என்றால் X என்பது Y-இன் வகையாகும் அல்லது இனமாகும்.

- பெயர்களுக்கு X என்பது Y என்பதன் வகையாகும்
- வினைகளுக்கு X என்பது Y என்பதன் ஒரு வழியாகும்

சினை-முழுமொழிகள் (meronomies)

சினைமுழுமொழிகள் சினை-முழு உறைவை விளக்கும். இது முழுமையுடன் பாகங்கள் காட்டுகின்ற வேறுபாட்டின் அளபுபையும் இப்பாகங்கள் பங்களிக்கின்ற பங்கேற்பாளர் கணக்கில் எடுக்க முயலும் கலவைத்தன்மையான உறவாகும். எடுத்துக்காட்டாக இடம்சார் இயைபும் (spatial cohesion) இட வகையீடும் (spatial differentiation) மற்றும் செயல்பாட்டு வகையீடும் (functional differentiation) பாகங்களுக்கு இடையே உள்ள தொடர்புகளின் இயல்பும் போன்ற கூறுகள் சினைமொழியங்களை நிர்ணயிக்கும் மிக முக்கியமான கூறுகளாகும். உண்மையில் இந்த கூறுகளின் தரம் அடிப்படையில் நாம் வேறுபட்ட பண்புகளைக்கொண்ட சினைமுழுமொழிகளின் வேறுபட்ட வகைகளைப் பெறலாம்.

சினைமுழுமைமொழிகளை மொழியியல் பரிசோதனைகளைப் பின்பற்றி பண்பாக்கம் செய்ய இயலும். பின்வரும் வாக்கியங்களில் ஒன்று சரி என்றால்தான் A என்பது B என்பதன் பகுதியாகும்.

B என்பது A ஐக் கொண்டிருக்கும் (அல்லது B ஒரு Aஐக் கொண்டிருக்கும்)

B என்பது Aயின் பாகமாகும்.

சினைமுழுமைமொழி உறவு எந்த உண்மை மாதிரியிலும் கருதப்பட வேண்டிய சில பண்புகளைக் (அல்லது அடைகளை) கொண்டிருக்கும்:

பாகத்தியின் விருப்பு

முழுமை அடிப்படையில் பாகத்தின் முக்கியத்துவம். எ.கா. மனிதனுக்கு இரண்டு கால்கள் உள்ளன; சிற்றுந்துக்கு நான்கு சக்கரங்கள் உள்ளன.

வின்ஸ்டன் சாபின் (Winston and Chaffin, 1987) செயல்பாடுகள் (functionalities), இட இயைபு (spatial cohesion), பகுதிகளுக்கும் முழுமைகளுக்கும் இடையிலுள்ள வேறுபாடுகளின் அளபு (degree of dissimilarity) என்பதன் அடிப்படையில் 6 வகையான சினைமுழுமைமொழிகளகளை வேற்படுத்துகின்றனர்; அவை பின்வரும் வகுப்புகளைத் தரும்:

உட்கூறு/ஒருங்கிணைந்த பொருள்: முழுமைக்கும் பகுதிகளுக்கும் இடையில் தெளிவான அமைப்புசார் மற்றும் செயல்பாடுசார் உறவு உள்ளது. எ.கா. பிடி/கோப்பை, ஒலியனியல்/மொழியியல்;

உறுப்பினர்/குழுமம் அல்லது குழு: பாகங்களுக்கு அமைப்புசார் அல்லது செயல்பாட்டு உறவு இருக்கவேண்டும் பாகங்கள் ஒன்றிலிருந்து ஒன்று தனித்துவமானதாக இருக்கவேண்டும் என்பதும் என்பது தேவையில்லை. இந்த வகுப்பில் மரங்கள்/காடு, மாணவர்கள்/வகுப்பு என்பன அடங்கும்.

பகுதி/திரள்: பகுதிகளுக்கு இடையிலும் பகுதிகளுக்கும் முழுமைக்கும் இடையிலும் முழு ஒற்றுமை இருப்பதில்லை. பகுதிகளுக்கு இடையிலுள்ள எல்லை/எல்லைகள் தன்னிச்சையானது; மற்றும் முழுமையைப் பொறுத்தவரையில் பாகங்களுக்கு முன்னோடியான குறிப்பிட்ட செயல்பாடு இல்லை. இந்த வகுப்பில் துண்டு/ரொட்டி, செண்டிமீட்டர்/மீட்டர். இந்த துணை உறவு சினை-முழுமை ஆய்வு என்று பெருப்பாலும் அழைக்கப்படும்.

பொருள்/மூலப்பொருள்: இந்த உறவு வகை ஒரு பொருளுக்கும் அது கட்டுமானம் அல்லது உருவாக்கம் செய்யப்படுகின்ற மூலப்பொருளுக்கும் அல்லது ஒரு பொருளுக்கும் அதன் ஆக்ககூறுகளுக்கும் உள்ள உறவை விளக்கும். எ.கா. மரம்/மேசை, தங்கம்/நகை.

துணைச்செயல்பாடு/செயல்பாடு அல்லது செயற்பாங்கு: ஒரு அமைப்பாக்க வழியில் ஒரு செயல்பாட்டை ஆக்கும் துணை செயல்பாடுகளை விளக்கும். எ.கா.கொடு/வாங்கு,

துல்லியமான இடம்/பரப்பு: பகுதிகள் முழுமைக்கு செயல்பாட்டு அடிப்படையில் உண்மையில் பங்களிப்பு செய்யாது. இத்துணை உறவு இடத்தை வெளிபடுத்தும். எ.கா. இமயமலை: இந்தியா.

வகைப்பாடுகளைப் போன்று சினைமுழுமொழிய உறவும் உண்மையில் இரண்டு கூறுகளுக்கு இடையில் மட்டும் கருத இயலாது; முழுமையின் எல்லாப் பாகங்களின் குழுமத்தின் அடிப்படையில் கருதப்படவேண்டும். இது ஒரு சினைமொழிய விளக்கத்தில் ஒரு வகையான அணுகுமுறையை அறிமுகப்படுத்த அனுமதிக்கும். சினைமுழுமொழியங்கள் பொதுவாக தருக்க நிலையிலும் மொழியியல் நிலையிலும் கடப்புத்திறத்தை (transitivity) அனுமதிக்காது. இருப்பினும் சில அறிஞர்கள் முன்னர் கூறிய சினைமொழிய உறவின் அதே துணைவகையால் இணைக்கப்பட்ட சொல்தனிமங்களுக்கிடையில் கடப்புத்திறத்தை அனுமதிக்கும் போக்கைக் காட்டுகின்றனர். இறுதியாக சினைமுழுமொழியங்களுக்கு முழுமையிலிருந்து பாகங்களுக்குப் பண்புகளின் கடப்புத்திறத்தின் மிகக்கூடுதலாகக் கட்டுப்படுத்தப்பட்ட வடிவம் இருக்கக்கூடும். இருப்பினும் இந்தக் கடப்புத்திறம் பொதுவான கொள்கைகளுக்குக் கீழ்படியாது மற்றும் ஒவ்வொரு விளக்கத்திலும் குறிப்பிடப்படவேண்டும். எடுத்துக்காட்டாக ஒரு சிற்றூர்தியியின் உடலுடன் தொடர்புடைய நிறப் பண்பு பொதுவாக முழுமையால், அதாவது சிற்றூர்தியால் மரபுரிமையாகப் பெறப்படுகின்றது. இந்த எடுத்துக்காட்டு ஓரளவுக்குப் பொதுவானதாகும்; இது மிதிவண்டிகள், வானவூர்திகள், தொடர்வண்டிகள், பேருந்துகள் எல்லாவற்றிற்கும் பொருந்தும். இருப்பினும் “முழுமை அதன் மிகப்பெரிய பாகத்தின் நிறத்தை மரபுரிமையாகப் பெறும்” என்ற பொதுவான கூற்று எப்பொழுதும் உண்மையல்ல.

கிளைகளல்லாத படிநிலைகள்

கிளைகளல்லாத படிநிலைகள் ஒழுங்கமைப்பு அல்லது பரிமாணத்தின் வேறுபட்ட மட்டங்களுடன் தொடர்புடைய சொல்தனிமங்களின் நிரலை அனுமதிக்கும். அமைப்பு வகை/துணைவகை ஒழுங்கமைப்புடன் பொருந்தாது; ஆனால் சில நேர்வுகளில் சினைமுழுமொழிய உறவுடன் ஒற்றுமை கொண்டிருக்கும். கிளைகளற்ற படிநிலைகள் பரிமாணத்தின் இடம்சார், காலம்சார் அல்லது அருவக் கருத்துச்சாயலுடன் உறவுகொண்டிருக்கும்.

நாம் கிளைகளற்ற படிநிலைகளின் மூன்று வகையை வேறுபடுத்தலாம்:

சொல்தனிமங்களுக்கிடையில் உள்ள எல்லைகள் தெளிவில்லாமல் இருக்கும் தொடர்ச்சியான படிநிலை; எடுத்துக்காட்டாக உறைநிலை – குளிர் – வெதுவெதுப்பு/நகச்சூடு – சூடு.

தொடர்ச்சியில்லாத மற்றும் தரப்படுத்த இயலாத படிநிலை; பொதுவாக நிர்வகப் படிநிலைகள் மற்றும் தொழில்நுட்பப் படிநிலைகள் போன்று ஏதாவது அளக்கவியலும் பண்பு அடிப்படையில் அமையாதவை: வாக்கியம் - தேற்றக்கூற்று - தொடர் – சொல் – உருபன்.

படிநிலையலலா உறவுகள்

படிநிலையலா உறவுகளில் கீழ் நாம் ஒருபொருள்பன்மொழியங்களையும் எதிர்ப்புகளில் வேறுபட்ட வடிவங்களையும் வேறுபடுத்துகின்றோம். இவ்வுறவுகள் ஈரிணையானதாகவோ மூவிணையானதாகவோ இருக்கும். மூவிணையான சிறப்பியல்பு இவ்வுறவுகள் சிலவற்றின் சூழல்சார்பைப் பிரதிபலிக்கும்.

ஒருபொருள்பன்மொழிகள்: இரு சொற்களுக்கு குறிப்பிடத்தக்க ஒற்றுமையான பொருண்மைப் பொருளடக்கம் இருந்தால் அவை பல்பொருளொருமொழிகள் ஆகும். ஒருபொருள்பன்மொழிகள் குறிப்பிடத்தக்க மேலுறலைக் கொண்டிருக்கும்; ஆனால் ஒருபொருள்பன்மொழிய அளவு அந்த மேலுறலுடன் தொடர்புடையதாக இருக்கத் தேவையில்லை. மொழியில் சில முழுமையான ஒருபொருள்பன்மொழிகள் இருக்கும்; ஆனால் சொற்கள் தரப்பட்ட சூழல்களில் ஒருபொருள்பன்மொழிகளாக இருக்கவியலும். எனவே நாம் நாம் ஒருபொருள்பன்மொழிய உறவை மூவிணை உறவு என்கின்றோம்: W1, W2 என்பன C என்ற சூழலில் ஒருபொருள்பன்மொழிகளாகும். ஒருபொருள்பன்மொழிகள் பெரும்பாலும் பொருண்மையியல் விளக்கங்களின் துல்லியத்தின் அளபைச் சார்ந்திருப்பதில்லை; ஒருபொருள்பன்மொழிகளின் அளபு நுணுக்கத்தின் வேறுபட்ட நிலைகளில் மாறும் (Spark Jones, 1986).

எடுத்துக்காட்டாக, கணினி அறிவியலில் சூழலுக்குள் directory, repertoire என்பனவற்றை எடுத்துக்கொண்டால் முந்தையது யுனிக்ஸ் உலகத்தையும் பிந்தையது மக்கிண்டோஷ் உலகத்தையும் சாரும் என்றாலும் அவை ஒருபொருள்பன்மொழிகள் ஆகும். இதேபோன்று கூடுதல் state என்பதன் மிக அருவத்தன்மையான சூழலில் remember, memorize என்பன ஒருபொருள்பன்மொழிகளாகும்.

ஒருபொருள்பன்மொழிகளின் கருத்துச்சாய்லைச் சீர்செய்ய சில அறிஞர்கள் ஒரு சொல்லுடன் தொடர்புடைய முக்கியமான பண்புக்கூறுகளையும் இரண்டாம் நிலைப் பண்புக்கூறுகளையும் வேறுபடுத்தும் போக்கைக் வெளிப்படுத்துகின்றனர். இந்த வேறுபாடு

நிச்சயமாக அனுபவவாதத்தால் தூண்டப்பட்டுள்ளது; இருப்பினும் இதை நடைமுறையில் பயன்படுத்துவது மிகக்கடினமாகும்.

ஒருபொருள்பன்மொழியம் என்பது ஒரு சொல்சார் அறிவு அடிப்படை (lexical knowledge base) ஆகும்; இது ஒருபொருள்பன்மொழிகள் எனக் கூறப்படும் இரு சொற்களுக்குத் தரப்படும் பண்புக்கூறு மதிப்பீடுகளைப் பற்றிய ஒருங்கிணைப்பு கட்டுப்பாட்டின் ஒரு வகைப் பாத்திரத்தை (role) செயல்படுத்தும். இது இயற்கைமொழி உருவாக்கத்தில் ஒரு ஆர்வமூட்டும் பாத்திரத்தைச் (role) செயல்படுத்தும்; இங்கு சொற்களை சில சூழல்களில் அவற்றின் ஒருபொருள்பன்மொழிகளால் இடம்பெயர்க்கலாம்.

எதிர்மொழிகளும் எதிர்ப்புகளும் (antonyms and opposites)

எதிர்மொழிகளும் (antonyms) எதிர்ப்புகளும் (opposites) ஏறக்குறைய தெளிவாக வரையறைவிளக்கம் செய்யப்பட்ட மிகப் பெரிய பல்வகைக் கருத்துச்சாயல்களை உட்படுத்தும். இதன் அடிப்படையான வரையறைவிளக்கம் W1, W2 என்பன அதிக அளவிலான பொதுப் பொருண்மையியல் சிறப்புப்பண்புகளையும் கொண்டிருந்து குறைந்தது ஒரு மிகத்தேவையான பொருண்மை பரிமாணத்தில் குறிப்பிடத்தக்க வழியில் வேறுபட்டால் அவை எதிர்மொழிகளும் எதிர்ப்புகளும் ஆகும். ஒருபொருள்பன்மொழிகளைப் போன்று எதிர்மொழிகளும் எதிர்ப்புகளும் சூழல் அடிப்படையிலானவை; அதனால் ஒரு மூவிணையான உறவை அறிமுகம் செய்கின்றது. சில சொற்பொருண்மைகளின் இணைகள் மூலமுன்வகை அடிப்படையில் (prototypically) பிறவற்றைவிட அதிக எதிர்ப்புகளாகும் என்பதால் எதிர்ப்புகளின் வேறுபட்ட அளபுகள் உள்ளன.

எடுத்துக்காட்டாக *தொடங்கு, தொடர், நில்* என்பன சூழல் அடிப்படையில் எதிர்ப்புகளாகும். இதுபோன்று *நல்ல, கெட்ட* என்பன பொதுவாக எதிர்மொழிகளாக ஏற்றுக்கொள்ளப்பட்டவை; *தாய், தந்தை* என்பவைகளுக்கு இடையே உள்ள எதிர்மையைவிட கூடுதல் மூலமுன்வகையானவை.

எதிர்மொழிகள் முழு கருத்துருக் களத்தை (conceptual domain) உட்படுத்தும் கருத்துரு வெளியை (Conceptual space) ஒன்றையொன்றை விலக்கிவைக்கும் தனிப்கூறுகளாக இருக்கத் தேவையில்லை. *நல்ல, கெட்ட* என்பவவற்றில் உள்ளது போன்று மேலுறலோ இடைவெளியோ சாத்தியமாகும்; ஏனென்றால் ஒன்று நல்லதாகவோ கெட்டதாகவோ இல்லை என்று கூறுவதோ அல்லது ஒன்று நல்லதும் கெட்டதும் ஆகும் என்று கூறுவதோ சாத்தியமாகும். எதிர்மொழிகளின் சிறப்பு வகுப்பு, முழுக் கருத்துரு வெளியை இரண்டு மேலுறல் செய்யாத தனிக்கூறுகளாக பிரிக்கும் நிரப்பிகளாகும் (complementaries). குரூசில் (Cruse 1986) தூண்டல்-விளைவு வகையின் உறவை

உருப்படுத்தம் செய்யும் ஊட்டாடிகளின் வகுப்பு போன்ற (interactives) (எ.கா. வேண்டு என்ற சூழலில் *அனுமதி-மறு*) நிரப்பிகளின் பல வகுப்புகள் வரையறை விளக்கம் செய்யப்பட்டுள்ளன

எதிர்ப்புகளுக்குள் மற்றொரு ஆர்வமூட்டும் வகுப்பு திசை எதிர்ப்புகளாகும் (directional opposites). அவை அடிப்படையான இடவியல்சார் அல்லது கருத்துசார் திசை எதிர்மைகளாகும் (directional oppositions). இந்த வகுப்பில் கருத்து அடிப்படையில் ஒத்தறி அடிப்படையில் எளிதான எடுத்துக்காட்டுகள் வரும்: *தொடங்கு – முடி, மேலெ – கீழே, இறங்கு – ஏறு*.

சொற்பொருண்மையியயல் அறிவு அடிப்படையில் (lexical semantics knowledge base) எதிர்ப்புகளின் பங்களிப்பை வரையறை விளக்கம் தருவது ஓரளவுக்குக் கடினமாகும். ஒருபொருள்பன்மொழிகளைப் போன்று எதிர்ப்புகளும் எதிர்மொழிகளும் நேர்மை கட்டுப்பாடுகளின் பங்களிப்பை (role of integrity constraints) செயல்படுத்தும். இயற்கை மொழி உருவாக்கத்தில் அவற்றின் பயன்பாடு; எடுத்துக்காட்டாக அதிக அளவிலான எதிர்மறைகளைத் தவிர்ப்பதை வெளிப்படையாகச் செய்வது ஓரளவுக்குக் கடினமாகும்; ஏனென்றால் எதிர்ப்புகள் அல்லது எதிர்மொழி இணைகளில் சொல்தனிமங்களின் துருவத்தன்மை பல பயன்வழியியல்சார் காரணிகளை ஊடாட்டம் செய்யலாம். எடுத்துக்காட்டாக நாம் *இந்த புத்தகம் எவ்வளவு விலை மிகுந்தது?* என்று கேட்கலாம், ஆனால் *இந்த புத்தகம் எவ்வளவு விலை மலிவானது?* என்று கேட்பதில்லை. இறுதியாக, இரண்டு சொல்தனிமங்கள் எந்த அளபில் எதிர்ப்புகள் அல்லது எதிர்மொழிகள் என்பதை வரையறை விளக்கம் செய்வதற்கு மொழியியல் சோதனைகள் அல்லது ஆய்வு நெறிமுறைகள் துல்லியமாக வரையறவிளக்கம் செய்யப்படவேண்டும்.

6.7.1.2.5. மூலப்பொருண்மையியல்கள் (ontologies)

மூலப்பொருண்மையியல்கள் சொற்பொருண்மையியலிலும் (lexical semantics) செயற்கை அறிவியலிலும் (artificial intelligence) பெரிய களமாகும்; அவற்றின் எல்லை தெளிவில்லாததாகும். அடிப்படையில், ஒரு மூலப்பொருண்மையியல் ஒரு குறிப்பிட்ட பொருட்புலத்திற்கு அடிப்படைத் தனிமங்கள், வேறுபட்ட கருத்துருக்கள் மற்றும் அவற்றின் தொடர்புள்ள மொழியல் மெய் உருவாக்கங்கள் இவற்றால் உருப்படுத்தம் செய்யும் நோக்கம் உள்ள ஒரு முறையான ஒழுங்குமுறை ஆகும். மூலப்பொருண்மையியலின் மிகப் பரந்த பார்வை அந்தக் களத்தைப் பற்றிய பட்டறிவு, அணி மற்றும் உருவக அறிவு போன்ற களஞ்சிய அறவின் பல வடிவுகளை உள்ளடக்கலாம்.

மூலப்பொருண்மையியல்சார் விவரணைகள் பல சூழல்களிலும் பல வகைப் பயன்பாடுகளிலும் பயன்படுத்தப்படுகின்றன. பயன்பாட்டைப் பொறுத்தவரையில் பல

ஒழுங்குமுறைகள் அவற்றின் தரத்தையும் மூலப்பொருண்மையியல்சார் அறிவு இயற்கை மொழியியலில், அறிவுள்ள பயன்பாட்டாளர் முன்முகம் மற்றும் ஆவணங்களின் தானியங்கு மீட்பு போன்ற பயன்பாடுகளுக்கு மிக முக்கியத்துவம் வாய்ந்ததாகும்.

மூலப்பொருண்மையியல் விளக்கங்கள் பல செயல்முறைகளிலும் பயன்பாடுகளின் பல வகுப்புகளிலும் பயன்படுத்தப்படுகின்றது. பயன்பாட்டைப் பொறுத்தவரையில் பல ஒழுங்குமுறைகள் அவற்றின் தரத்தையும் பொதுத்தன்மையையும் மேம்படச்செய்ய மூலப்பொருண்மையியலைப் அறிவைப் பயன்படுத்துகின்றது. மூலப்பொருண்மையியல் அறிவு (ontological knowledge) இயற்கைமொழி மற்றும் ஆணவங்களின் தானியக்க மீட்பு இவற்றின் அறிவுத்திறனவாய்ந்த பயன்பாட்டாளர் முகப்பு முனைகள் (intelligent user-front ends) போன்ற பயன்பாடுகளுக்கு மிக முக்கியத்துவம் வாய்ந்ததாகும். நடைமுறை மற்றும் கோட்பாடு சார்ந்த அணுகுமுறைகளிலும் உட்படையாகப் பயன்படுத்தப்படுகின்றது: பொருண்மைசார் வலைகள் (semantic net works) மற்றும் கருத்துரு வரைபடங்கள் (conceptual graph), சொல் கருத்துரு அமைப்புகள் (lexical conceptual structures), ஆக்கமுறை அகராதிகள் (generative lexicons), காட்சிகள் (scenarios), எழுத்துகள் (scripts), கருத்தாடல் மேலாண்மை ஒழுங்குமுறை (discourse management system). இடம், காலம் போன்ற சில பொதுவான களங்கள் மிக விரிவான மூலப்பொருண்மையியல் ஒழுங்குமுறைகளை உருவாக்கின.

மூலப்பொருள் ஆய்வை விளக்க வரையறை செய்வது சிரமான ஒன்றாகும். ஒரு பொருண்மைக்களத்துடன் தொடர்புடைய உண்மையான அறிவின் மூலமுதல் அல்லது அடிப்படைக் தனிமங்களை/கூறுகளை வரையறைவிளக்கம் செய்ய இயலுவது தான் முக்கியமான விஷயம். இந்த விளக்க வரையறை விளக்கம் பல காரணிகளை அடிப்படையாகக் கொண்டது. அவற்றில் சில கீழே தரப்பட்டுள்ளன (Saint-Dizoer and Viegas19951995:20):

• ஒழுங்கு முறையில் எந்த அளவுக்கு நுண்மையை ஒருவர் விரும்புகிறார் என்பதன் தரம்.

• ஏற்கனவே இருக்கிற இட மற்றும் கால மூலப்பொருண்மையியல்களை ஒழுங்குமுறையில் சீராக உட்படுத்துவதைக் குறிப்பிடுதல்.

• சொல் கருத்துரு அமைப்பு (Lexical Conceptual Structure) (LCS) (Jakendoff, 1990) அல்லது கருத்துரு வரைபடச் சட்டகம் (Conceptual Graph Framework) போன்ற மொழியியல்சார் அறிவை உருப்படுத்ததின் (linguistic knowledge representation) கோட்பாடுகளை குறிப்பிடுதல்.

ஒரு மூலப்பெருண்மையியல் பின்வருபவைகளை உள்ளடக்கும் ஒரு முறையான மொழி அடிப்படையில் அமைந்ததாகும் (Saint-Dizoer and Viegas 1995:20):

• பெரும்பாலும் அமைப்பாக்கம் செய்யப்பட்ட மற்றும் வகைப்படுத்தப்பட்ட இருப்புப்பொருள்களின் (entities) குழுமம்.

• ஒரு குழும உறவுகள் மற்றும் செயல்பாடுகள் மற்றும் அவற்றின் தொடர்புள்ள பண்புகள்.

• ஒரு பொருண்மைக்களத்தின் மூலமுதல் நிலைகளையும் செயல்களையும் விவரணை செய்யும் அடிப்படையான பயனிலைகளின் முற்றுக் குழுமம்; அவற்றின் முறையான பொருண்மையில் முறையே அவற்றின் பயன்பாட்டுக்குட்பட்ட பொருள்கோளின் களங்களால் விவரணைசெய்யபடுகிறது.

• இருப்புப்பொருள்களின் மேல் செயல்படும் செயல்பாடுகளின் குழுமம். இச்செயல்பாடுகள் எல்லாம் தரவுத்தொகுதி ஆய்வால் (corpus analysis) வரையறை விளக்கம் செய்யப்படுகின்றன.

இந்த அறிவு முதல் கிரமத் தர்க்கத்தின் துணைக் குழும் ஆகும். இது பொதுமையின் பல மட்டங்களில் மொழியில் மற்றும் பயன்பாடுசார் அறிவை உருப்படுத்தம் செய்ய பயன்படுத்தப்படுகின்றது. கீழ் வருவது எளிதான வங்கிக் களத்தின் (bank domain) மூலப்பொருண்மையியலாகும் (Saint-Dizoer and Viegas, 1995:20):

(17) I: 2 துணைவகைகளையும் உள்ளடக்கிய தனிநபர்கள் மனிதவினம் (துணைவகைப்பாடு செய்யப்பட்டுள்ளது: வாடிக்கையாளர் மற்றும் வேலையாள்) நிறுவனங்கள் (துணைவகைப்பாடு செய்யப்பட்டுள்ளது: வங்கி, தொழில் நிறுவனங்கள் மற்றும் அரசு மற்றும் தனியார் நிறுவனங்கள்)

A: பணத்தின் பருமைத் தொகைகள். இரண்டு தசமங்கள் உள்ள வேரளவையல்லா எண்கள் உள்ள வகைப்;பாட்டு இடம் முழுமையான நிரல் உறவு உள்ளது <, சமன்பாடு மற்றும் கணிதச் செயல்பாடுகள்.

M: காகிதம், பிளாஸ்டிக் மற்றும் உலோகம் போன்ற பருமைப் பொருள்கள்.

S: அருவத்தன்மையுள்ள துணைநிற்பவைகள், எ.கா. காசோலை மற்றும் சில்லறைகள் பணத்திற்குத் துணைநிற்பவைகள்.

T. நிகழ்வுகளின் குழுமம். மேற்கண்ட சொல் கருத்துரு அமைப்பு அடிப்படையில் பணத்தின் சலனத்தை பின்வருமாறு உருப்படுத்தம் செய்யலாம்.

(18) சலனம்(கையிருப்பு_இன், i_1, i_2, தொகை) = உரியது_க்கு (i_1 தொகை t_1)^உரியது_க்கு (i_2, தொகை, t_2)^வேறுபாடு (i_1,i_2)^ t_2>t)

6.7.1.2.6. சொல்சார் கருத்துரு அமைப்பு (Lexical Conceptual Structure (LCS))

சொல் கருத்துரு அமைப்பு (Jakendoff, 1988, 1990) மூல கருத்துருச் சாயல் அடிப்படையில் வெளி மொழிக்கும் அகமொழிக்கும் இடையே முறையான பொருத்தை நிறுவுவதை நோக்கமாகக்

கொள்கின்றது. அடிப்படைக் கருத்துருக்கள் சொல்சார்ந்தவை மற்றும் வாக்கியத்தின் பொருண்மை சொற் பொருண்மையால் உருவாக்கப்படுகின்றது. LCS பின்வரும் அடிப்படையான எடுகோள்களை நிறுவுகின்றது (Saint-Dizoer and Viegas, 1995:21).

• எந்தத் தொடரியல் உறுப்புக்கும் கருத்துரு உருப்படுத்தம் உண்டு.

• எந்தக் கருத்துரு வகைப்பாடும் ஒன்றோ அதற்குமேற்பட்ட மொழியியல் அலகுகளை குறியாக்கம் செய்கின்றது.

• பெரும்பாலான இந்த வகைப்பாடுகள் வகைக்கும் பொருளுக்கும் (சாதனத்திற்கும்) வேறுபாடு பாராட்டுகின்றது.

• இந்த வகைப்பாடுகளை அளவையாக்கம் செய்ய இயலும்.

• ஒரு சொல்லின் கருத்துரு அமைப்பு பூஜிய அல்லது அதற்கு மேற்பட்ட பங்கெடுப்பாளர்கள் உடன் கூடிய ஒரு உள்பொருளாகும்.

6.7.1.2.7. நிகழ்வு அமைப்புகள் மற்றும் காலச் சொல்சார் தகவல் (event Structure and temporal lexical information)

அமைப்புகள் ஒரு சொல்லோடோ தொடரோடோ தொடர்புடைய ஒரு குறிப்பிட்ட வகையிற்படும் நிகழ்வை அடையாளம் காண அனுமதிக்கின்றது (Vendler, 1967). அடிப்படையில், ஒரு சொல்லோ தொடரோ ஒரு நிலைமையையோ (ளவயவந) ஒரு செயற்பாங்கையோ (process) ஒரு மாற்றநிலையையோ (transition) குறிப்பிடுகின்றது. ஒரு நிலை என்பது எந்த செயலுடனும் ஒப்பீட்டு அடிப்படையில் முழுவதும் மதிப்பீடு செய்யப்படாததாகும். பின்வரும் சொற்கள் நிலைகளைக் குறிப்பிடும் (Saint-Dizoer and Viegas 24): உறங்கு, நம்பு, வெண்மை.

ஒரு செயற்பாங்கை ஒரு தொடர்ச்சியான ஒரேமாதிரியான செயல்களாக ஆயலாம். எடுத்துக்காட்டாக ஓடுதல் என்ற செயலை சிறிய தூரம் ஓடுதலுக்குச் சமமான (எல்லையற்றத்) துணை செயல்களாகச் சிதைக்கலாம். இறுதி எல்லையில் இந்தச் செயலை ஓடுதலுடன் தொடர்புடைய அடிப்படைச் சைகைகளாகச் சிதைக்கலாம். இறுதியாக மாற்றநிலைகள் என்பன ஒரு எதிர்மறையுடன் தொடர்புடைய பயனிலைகளாகும். மாற்றநிலைகள் ஒரு செயற்பாங்கால் ஒரு நிலையிலிருந்து மற்றொரு நிலைக்கு மாறுவதை விளக்கும் (Pustejovsky, 1991, 'The syntax of event structure, Cognition'). மாற்றநிலையை திற என்ற வினையால் விளக்கலாம். அடிப்படையான கருத்து என்னவென்றால் கதவைத் திறப்பதற்கு முன் கதவு அடைத்திருக்கும். திறக்கிற செயல் நிகழ்த்தப்படும்போது, கதவு திறக்கிற துணைநிலையை அடையும்.

6.7.1.2.8. குண மற்றும் மரபுரிமை அமைப்புகள் (The qualia and inheritance structure)

ஜேம்ஸ் புஸ்தோவ்ஸ்கி (James Pustejovsky, 1995) மிக முறைபடுத்தப்பட்ட அமைபப்பாக்கம் செய்யப்பட்ட ஆக்கமுறை அகராதியை உருவாக்குகிறார் (Saint-Dizoer and Viegas, 1995:25).

• பங்கேற்பாளர் அமைப்பு (Argument Structure): இது ஒரு சொல்லின் பயனியலை அமைப்பை உருப்படுத்தம் செய்கின்றது.

• நிகழ்வு அமைப்பு (Event Structure): (நிலை, செயற்பாங்கு அல்லது மாற்றம் என்ற) மூன்று வேறுபட்ட நிகழ்வுவகுப்புகளும் உள்ளடங்கிய குறிப்பிட்ட நிகழ்வு வகையைக் கண்டுபிடிக்கின்றது.

• பண்பு அமைப்பு (Qualia Structure): இது பின்வரும் பாத்திரங்கள் மூலம் கொள்ளப்பட்ட சொற் பொருண்மையின் நான்கு விஷயங்களுக்கிடையே வினியோகிக்கப்பட்ட ஒருபொருளின்/சாதனத்தின் முக்கியமான அடைமொழிகளைத் தருகின்றது.

- ஒரு பெரிய களத்திற்குள் அதனை வேறுபடுத்திக் காண்பிக்கின்ற முறையான பங்களிப்பு (formal role)
- சொல்லுக்கும் அதன் உறுப்புகளுக்கும் அல்லது சரியான பாகங்களுக்கும் இடையேயுள்ள உறவைத் தரும் ஆக்கக்கூறு பங்களிப்பு (constitutive role), அதன் தேவையையும் இலக்கையும் விளக்கும் நோக்கப்பங்களிப்பு (telic role)
- எது அதை வெளிக்கொணர்கின்றது என்ற செய்தியைத் தருகின்ற செயலி பங்களிப்பு (agent role)

• மரபுரிமை அமைப்பு (Inheritance structure): இது இரு அறிவுபூர்வ நுட்பத்தை உட்படுத்தும்.

• பண்பு அமைப்பால் வலுவூடப்பட்ட உறவும் மரபுசார்ந்தது என்றும் கூறப்படுகின்ற நிரந்தர மரபுரிமை (fixed inheritance).

• ஒரு கருத்துருவுக்கு இறுதி இடம்தருகிற அதாவது பட்டறிவுடன் தொடர்புடைது என்று பெரும்பாலும் கூறப்படுகின்ற தனிப்பட்ட வகைப்பாடுகளுக்கு,

கருத்துருக்களுக்கு அதன் உறவமைப்பு உருவாக்குகிற முற்கொணர் மரபுரிமை (projection inheritance) பின்வருவது புத்தகம், அகராதி என்பதன் பண்பு அமைப்புகளாகும்:

(19) புத்தகம் (*x *)

[முறையான: புத்தக வடிவம் (*x *)]

[ஆக்கக்கூறு: {பக்கங்கள், அட்டை,...}]

[நோக்கம்: படி (T,y,*x*)]

[செயலி:உருவாக்கப்பட்ட பொருள் (*x *), எழுது (T,z,*x*)]

(20) அகராதி

[முறையான: புத்தகம், மின்தட்டு (*x *)]

[ஆக்கக்கூறு: அகரவரிசைப்படுத்தப்பட்ட பட்டியல் (*x *)]

நோக்கம்: நோக்கீடு (P,y,*x*), தகவல் தேடுதல் (T,y,*x*)]

[செயலி:உருவாக்கப்பட்ட பொருள் (*x *), தொகு (T,z,*x*)]

(21) நாவல் ஒரு புத்தகம் அகராதி ஒரு புத்தகம்

படி	ஆம்	?
வாங்கு	ஆம்	ஆம்
தகவல்தேடு	?	ஆம்

6.7.1.2.9. பொருண்மை-பனுவல் கோட்பாடு சொல்சார் செல்பாடுகள் (Lexical functions in meaning-text theory)

மெல்சுக் (Mel'čuk, 1988) எனபொருண்மை-பனுவல் கோட்பாடு (Meaning-text theory =MTT)) கணினிமொழியியலில் எடுத்துக்கொள்ளப்பட்டுள்ளது. பொருண்மை-பனுவல் கோட்பாடு பனுவல்களை அவற்றின் பொருண்மை உருப்படுதத்துடன் தொடர்புபடுத்தி ஒரு குழும இயல் மாதிரிகளை (பொருண்மை-பனுவல் மாதிரிகளை (பொ.ப.மா)) விளக்குகின்றது. பொ.ப.மா. இயற்கை மொழி பனுவல்களுக்கும் பொருண்மைகளுக்கும் இடையே உள்ள ஒப்புவரைவை விளக்குகின்றது (Saint-Dizoer and Viegas.1995:27).

1. பொருண்மை உருப்படுத்தங்கள் (semantic representations (SemR))
2. அக தொடரியல் உருப்படுத்தங்கள் (deep syntactic representations (DSyntR))
3. புற தொரியல் உருப்படுத்தங்கள் (surface syntactic representations SSyntR)
4. அக உருபனியல் உருப்படுத்தங்கள் (deep mophological representations (DMorphR))
5. புற உருபனியல் உருப்படுத்தங்கள் ((surface mophological representations (SMorphR))
6. அக ஒலியனியல் உருப்படுத்தங்கள் (deep phonetic representations (DphonR))
7. புற ஒலியனியல் உருப்படுத்தங்கள் (surface phonetic representations (SPhonR))

பொ.ப.மா.-வில் உள்ள செய்திகள் விளக்க இணைப்பு அகராதியில் (Explanatory Combinatorial Dictionary (ECD)) தரப்பட்டுள்ளது. ECD எல்லா மொழியியல் அறிவையும் (உலகஅறிவுக்கு எதிரான) உள்ளடக்க முயற்சிக்கின்றது. சொல் பதிவுகள் பொருண்மையியல் மண்டலம் (semantic

zone), தொடரியல் மண்டலம் (syntactic zone), சொல் இணைப்பு மண்டலம் (lexical combinatorial zone) என்ற மூன்று மண்டலங்களாகப் பிரிக்கப்பட்டுள்ளன.

• பொருண்மை மண்டலம் எளிமையான பொருண்மைத்தன்மைவாய்ந்த கூறுகளால் சொல் பதிவின் பொருண்மையை விளக்கும் பொருண்மைப் பின்னலை குறிப்பிடுகின்றது. அதாவது ஒரு சொல்லின் பொருண்மை மிகக்குறைந்த கலப்படமான சொல்நடையால் பெயர்புரை செய்யப்பட்டுள்ளது. எடுத்துக்காட்டாக விலக்கு என்பது X Y-ஐ விலக்கு என்று விவரணை செய்யப்பட்டுள்ளது.

• தொடரியல் மண்டலம் சொல்பதிவின் தொடரியல் வகுப்பையும் சொல்லனை ஏற்கும் குறிப்பிட்ட கட்டமைப்புகளை அடையாளம் காணும் தொடரியல் பண்புக்கூறுகளையும் எவ்வாறு ஒரு சொல்பதிவின் பொருண்மை செயலிகள் அகத்தொடரியல் உருப்படுத்தம் மற்றும் புறத்தொடரியல் உருப்படுத்தம் என்ற இரு தொடரியல் மட்டத்திலும் வெளிப்படுத்தப்படுகின்றது என்பதைக் காட்டுகின்ற ஆளுகை அமைப்பொழுங்கையும் குறிப்பிடுகின்றது.

• சொல் இணைப்பு மண்டலம் பொருண்மையால் உறவுபடுத்தப்பட்ட சொல்லன்களை சொல் செயல்பாடுகளின் (Lexical functions (LFs)) மதிப்பாக குறிப்பிடுகின்றது.

F என்ற சொல் செயல்பாடு F -இன் முக்கிய சொல் என்று அழைக்கப்படுகின்ற L என்ற சொல்லை ஒரு குழும சொற்களுடன் F(L) தொடர்புபடுத்துகின்ற பொருத்தமாகும். இரண்டு வகையிலான சொல் செயல்பாடுகள் உள்ளன.

• அடுக்குறவு சொல் செயல்பாடுகள் (paradigmatic LFs): இவை ஒருபொருள்பன்மொழிகள், உள்ளடக்குச்சொற்கள், மறதலைச் சொற்கள், இவற்றை கணிக்கின்றன. இவை ஒரே தொடரியல் வகுப்பைச்சார்ந்த பொருண்மை அடிப்படையில் தொடர்புள்ள சொற்களை அடையாளம் காட்டுகின்றது. எடுத்துக்காட்டாக Syn (விலக்கு):தடைசெய்.

• உறுப்பமைவு சொல் செயல்பாடுகள் (syntagmatic LFs): இவை ஒரே வாக்கியத்தில் பதிவுச்சொற்களுடன் குறிப்பிட்ட தொடரியல் உறவுகளில் சோ;ந்துவருகின்ற சொற்களின் (உடிடடடிஉயவiடிஇயட றடிசனள) மதிப்பைக் கணிக்கின்றது.

6.7.2. மன அகராதியும் இயந்திர அகராதியும் (Mental lexicon and machine lexicon)

ஜீன் பிரான்சிஸ் லே நி (Jean-Francois Le Ny, 1995:50-67) மன அகராதியையும்; இயந்திர அகராதியையும் ஒப்பிட்டுப் பேசுகிறார். மொழி புரிந்துகொள்ளுதலும் அர்த்தங்கொள்ளுதலும் (comprehension and interpretation) மனித உள்ளங்களையும் (human minds) அறிவுபூர்வமான பொறி அமைப்பொழுங்களையும் (intelligent machine sytems) உள்ளடக்குகின்றது. பொதுப்

புலனறிவு அறிவியல் பார்வையில் 'எப்பண்புகள் இயந்திரங்களாலும் உளச்சொல் உருப்படுத்தத்தாலும் பங்கிடப்படுகின்றது? எவை பங்கிடப்படவில்லை?' என்பது கேள்வியாகும்.

6.7.2.1. இயற்கை மற்றும் செயற்கை பொருண்மை அலகுளின் பொதுவான பண்புகள்

கணினி பொருண்மையியலில் பொருண்மை அலகுகள் அல்லது புலனறிவு உளவியலில் அறிவு உருப்படுத்தம் இவற்றுடன் அடைப்படுத்தப்பட்ட குணங்களின் குழுமங்கள் பட்டியலிப்பட்டுள்ளன. இப்பட்டியல் மூன்று பாகங்களாக பிரிக்கப்பட்டுள்ளன. முதல் இருபாகங்கள் மிக அறியப்பட்ட பண்புகளையும் மூன்றாவது பாகம் மனஉருப்படுத்தங்களின் கூடுதலான பண்புகளையும் உள்ளடக்கி இருக்கின்றது (Jean-Francois Le Ny, 1995:52).

அ. அடிப்படை பண்புகள்

1. குறிப்பீடு அல்லது வகைப்பாடு (Denotation or categorization)
2. உள்ளடக்கல்/உள்ளடங்கல் (Superordination/Subordination)
3. அடைப்படுத்துகை (Attribution)
4. வேற்றுமை பங்களிப்புகளில் இறுப்பானின் தேவை அல்லது வினை அல்லது வின பெயராக்கங்களின் செயபடுத்தம் (Filler requirement in case rolses or actanciation of verbs or verb nominalizations)
5. பொருண்மை அல்லது கருத்துரு அமைப்புகள் (Semantic or conceptual structures)

ஆ. இரண்டாம் நிரல், ஆக்கப் பண்புகள் (Second order, derived properties)

6. உட்படுத்தல் மரபுரிமைக் கட்டுப்பாடுகள் (Entailments)
7. மரபுரிமை (Inheritance)

இ. மிகத் தெரிந்த உளவியல் பண்பகள் (Well know phsyhological properties)

8. ஒற்றுமை (Similarity)
9. தனித்தேகத் தன்மை (Typicality (representativity))
10. படிநிலையில் அடிப்படை மட்டங்கள் (Basic levels in hierarchies)

6.7.2.2. நடைமுறைப்படுத்தும் நுட்பத்தில் உள்ள அடிப்படை வேறுபாடுகள் (A basic difference in the processing mechanism)

புலனறிவு உளமொழியியலில் தற்போது மொழிப்புரிந்துகொள்கையின் மாதிரியில் இரு அகன்ற வகுப்புகள் பயன்படுத்தப்படுகின்றன. அவை நடைமுறைப்படுத்தும் நுட்பங்களின் இரு வேறுபட்ட பார்வைகளை அடிப்படையாகக் கொண்டன: முதலாவது கணினி உருவகத்தை (computer metaphor) சார்ந்திருப்பது. இரண்டாவது நரம்பு உயிரியலுக்கு (neurobiology)

அண்மைப்படுவது. இரண்டும் புரிந்துகொள்கையைப் பொறுத்தவரையில் ஒரே பொதுப்பார்வையைக் கொண்டிருக்கின்றன. இந்த இரு உளவியல் மாதிரிகளின் இரு வகுப்புகளுக்கு இடையிலான முக்கியமான வேறுபாடு பொருண்மை உருப்படுத்தத்தின் தன்மை மற்றும் இடத்தில் இருக்கின்றது.

மேலும் பின்வரும் கருத்துக்களிலும் மன அகராதியும் இயந்திர அகராதியும் வேறுபடுகின்றன (Jean-Francois Le Ny, 1995):

• அகராதி அலகுகளின் குறிப்பிடும்படியான பண்புகள் (Specific properties in the units in lexicon)

• மனித மற்றும் இயந்திரத்தில் தூண்டுகை மற்றும் அணுகுகை: மயக்கநீக்கம் (Activability and accessibility in man and machine: Disambiguation)

• சூழல் தூண்டுகை மற்றும் முன்தூண்டுதல் (Contextual activability and pre-activiation)

• பொருண்மை பண்புக்கூறுகளை தூண்டுகை (Activability of semantic features)

6.7.3.. சொல் பொருண்மையியலிலிருந்து பனுவல் ஆய்வு (From Lexical Semantics To Text Analysis)

சபைன் பெர்க்லர் (Sabine Bergler, 1995: 98-124) சொற்பொருண்மையியலிலிருந்து பனுவல் ஆய்வுக்குச் செல்வதை பின்வருமாறு விளக்குகிறார். இன்றைய முக்கியமான அறைகூவல்களில் ஒன்று சாத்தியமான தகவல்களின் அளவுக்கு அதிகமான தன்மை. பின்வரும் பகுதிகள் முக்கியமானவைகள் ஆகும்:

• ஊக்கம் (Motivation)

• பத்திரிகைப் பனுவலில் கூற்றுப் பேச்சு (Reported Speech in newspaper texts)

• வாதிடுகை அமைப்பு (Argumentative structure)

• மூலம் (Source)

• சொற்பொருண்மையியல் (Lexical Semantics)

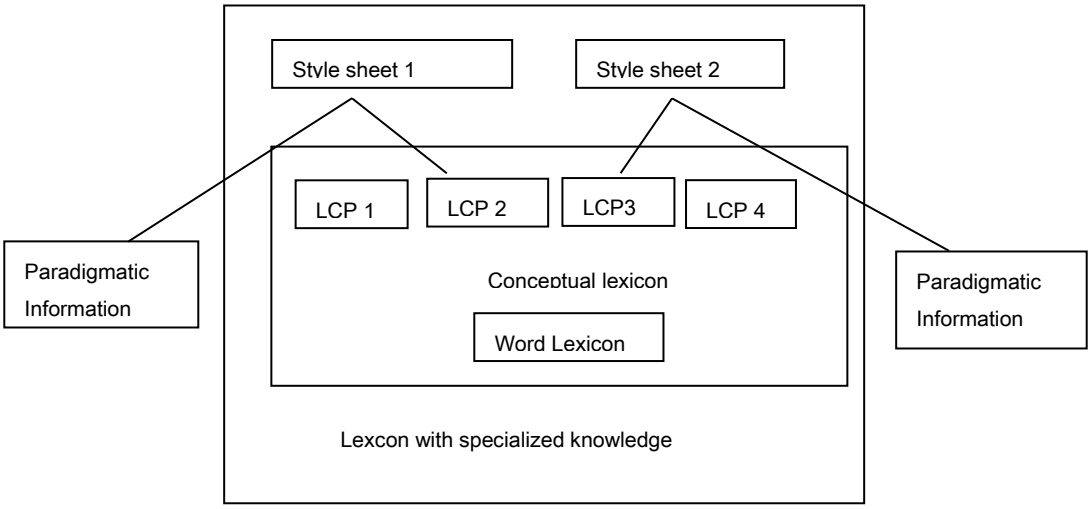
• அடுக்கு அகராதி திட்டமுன்மாதிரி (Layered lexicon design)

• மூல உருப்படுத்தம் (Source Representation)

- அமைப்பு (Profiles)
- உதவும் குழுக்கள் (Supporting groups)

• துணைபுரியும் குழு அமைப்புகளை ஆக்குதல் (Deriving supporting group structures) அடுக்கு அகராதி திட்டவரைவைப் (Layered Lexicon Design) பின்வருவமாறு வரைபடமாகக் காட்டலாம் (Jean-Francois Le Ny, 1995: 108).

(22)



அகராதியில் *official*, *agent* என்பதன் விவரணை பின்வருமாறு அமைகிறது.

(23) official n.1. One who holds an office or position

office3. A positon of authority given to a person, as in government or other organizaton.

agent n 1.One that acts or has the power to act.

2.One that acts as the representative of another

சொல்பதிவுகள் (Lexical Entries) பின்வருமாறு அமையும்.

(24)

official (x,y)
CONST: individual (x)
FORMAL: human (x), organization (y)
position (z), in (z,y), holds (x,z),
has – authority (z), trusty (y,z)
TELIC:IN (z)
AGENIVE:INH (y),
defalt: hire (y,x)

(25)

agent (x,y)
CONST: individual (x)
FORMAL: human (x)
TELIC: represent (x,y)
AGENTIVE: inh (Y)
default: hire (y,x)

பத்திரிக்கை கட்டுரைகளில் உள்ள கூற்றுப்பேச்சுக்குத் தேவையான உருப்படுத்தம் தர இயலும் என்பதை சபைன் பெர்க்லர் நிறுவுகிறார்.

2.5. சொல் செயல்பாடுகள், ஆக்கமுறை அகராதிகள் மற்றும் உலகம் (Lexical Functions, generative lexicons and the world)

டிர்க் ஹெலைன் (Dirk Helen, 1995:125-140) சொல் செயல்பாடுகள், ஆக்கமுறை அகராதிகள் மற்றும் உலகம் இவற்றிற்கு இடையே உள்ள உறவை பின்வருமாறு விளக்குகிறார். இயற்கைமொழிப் பொருண்மையியலுக்கும் குறிப்பாக சொற்பொருண்மையியலுக்கும் அறிவு உருப்படுத்தத்திற்கும் (knowledge representation) இடையில் உள்ள தொர்புகள் பலவகைப்பட்டது. சொற்பொருண்மையியல் இந்த நிலையைப் பிடித்துக்கொண்டிருப்பதற்கும் காரணத்தைப் பொருண்மையின் கூறு குறிப்பீட்டுக் கோட்பாடைப் (compositional denotational theories of meaning) பார்த்தால் தெரியும். இங்கு வெளிப்பாடுகளின் பொருண்மையை மொழி

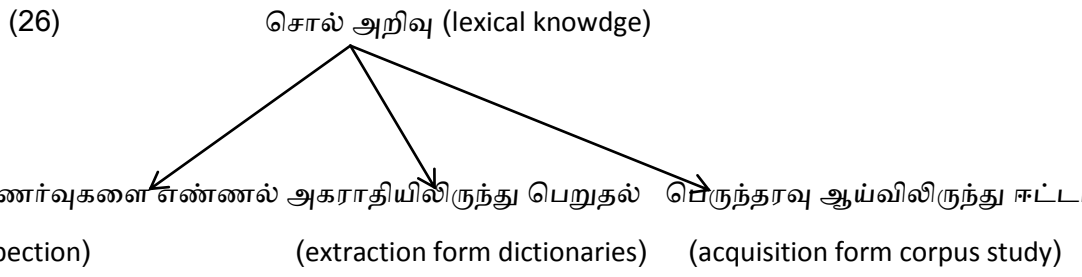
வெளிப்பாடுகளுக்கும் உலகத்திற்கும் இடையேயுள்ள உறவால் குறிப்பிட முயலுகின்றனர். அகராதி ஒவ்வொரு அடிப்படையான சொல்லும் உலகின் எந்தப் பகுதியைக் குறிப்பிடுகிறது என்பதை வெளிப்படுத்துகின்றது. ஆனால் இலக்கண விதிகள் பகுதிகளின் பொருண்மைகளை முழுமைகளின் பொருண்மைகளுடன் சோ;க்கும் பொதுவான கட்டளைகளுடன் தொடர்புடையதாய் இருக்கின்றது. உண்மையான விளக்க அகராதிகள் இயற்கைமொழி வெளிப்பாடுகளுக்கிடையே உள்ள கலவைத்தன்மையான உறவுகளை கூறுவதாகக் கருதலாம். அறிவு உருப்படுத்தச் சிக்கல்களைச் சொற்பொருண்மைச் சிக்கல்களுடன் முழுவதுமாகத் தொடர்புபடுத்த இயலாது என்றாலும் பின்னையதைக் கூறுகையில் பின்னையதைப் பார்ப்து நல்லது என்பதற்குத் தகுந்த காரணங்கள் இருக்கிறது.

6.7.4. சொற்பொருண்மையியல்: அகராதியா அல்லது களஞ்சியமா? (**Lexical Semantics: Dictionary Or Encyclopaedia?**)

மார்க் கவாசா மற்றும் பியரே வெய்கன்பாம் (Marc Cavazza and Pierre Zweigenbaum, 1995, 336-347) என்பவர்கள் பொருண்மை அகராதி (semantic lexion) ஒரு அகராதி வடிவில் அமையுமா அல்லது களஞ்சிய வடிவில் அமையுமா என்ற கேள்வியை எழுப்பி சொற்பொருண்மையியலுடன் தொடர்புபடுத்தி பின்வருமாறு விளக்குகிறார்கள். இயற்கை மொழியைப் புரிந்து கொள்ளும் அமைப்பொழுங்கில் உள்ள அடிப்படையான விசயங்கள் வாக்கிய அமைப்பை நிர்ணயிக்கும் தொடரியல் பகுப்பாய்வி (syntactic parser) மற்றும் இயற்கை மொழி உள்ளீட்டிலிருந்து அமைப்பொழுங்கின் உண்மை அறிவை அணுகப் பயன்படும் பொருண்மை அகராதி (semantic lexion) என்பனவாகும். பொருண்மை அகராதி சொற்களை உலக அறிவுடன் தொடர்புபடுத்துவதில் முக்கியமான பங்கு வகிக்கின்றது. ஆனால் பொருண்மை அகராதி வாக்கிய அமைப்பை நிர்ணயிக்கையில் குறிப்பிட்ட மொழியியல் சிக்கல்களுக்குத் தீர்வுகாணப் பயன்படுத்தப்படுகின்றது. அகராதி-களஞ்சிய முரண்பாடு சொல்விளக்கத்தின் இரு தீவிரமான கருத்துக்களை எதிர்க்கிறது: அகராதி அணுகுமுறை அடிப்படையில் ஒரு சொல் உலக அறிவோடு தொடர்புபடுத்தப்படாமல் மொழிச்சொற்களால் தான் விளக்கபடவேண்டும். களஞ்சிய விவரணை சொல்லால் குறிப்பிடப்படும் வேறுபட்ட இனங்கள் அல்லது ஒரு பொருளின் வேறுபட்ட நிலைகள், இந்தப் பொருளின் முக்கிய நடத்ததை வகைகள் அல்லது செயற்பாங்கு இவற்றை உள்ளடக்கும்.

பின்வரும் படம் சொல் செய்தி ஈட்டத்தின் (lexical information acqusition) அணுகுமுறைகளை சுருக்கமாகக் கூறுகிறது.

(26)



கணினி சொற் பொருண்மையியலின் முக்கிய வாதம் பெருந்தரவிலிருந்து மொழி அறிவை ஈட்டுவதாகும்.

6.8. சுருக்கவுரை

இவ்வியலின் தொடக்கத்தில் கணினி பொருண்மையியல் குறித்த ஒரு சுருக்கமான அறிமுகம் செய்யப்பட்டுள்ளது. அதன் பின்னர் கணினி சொற்பொருண்மையியல் குறித்து விளக்கம் தரப்பட்டுள்ளது. கணினியைப் பயன்படுத்திச் சொற்பொருண்மை ஆய்வை மேற்கொள்ளும் செயல்பாடுகள் கணினி சொற்பொருண்மையியல் என்ற புதிய பாடக்கிளையை உருவாகும் அளவுக்கு வளர்ந்துள்ளது. சொற்பொருண்மை ஆய்வு என்றால் கணினியைப் பயன்படுத்திச் செய்யப்படும் ஆய்வு என்று கூறுகிற அளவிற்கு இக்கிளை இயற்கை மொழி ஆய்வில் கால் ஊன்றியுள்ளது. இங்கு கூறப்பட்ட செய்திகள் கணினி சொற்பொருண்மையியலின் பல பரிமாணங்களை அறியத் துணைபுரியும். ஆக்கமுறை அகராதி ஆய்வும் கணினி சொற்பொருண்மையியல் ஆய்வின் ஒரு வகையாகும்.

Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

இயல் 7

எழுத்துப்பிழை திருத்தமும் இலக்கணப்பிழை திருத்தமும்

7.1. அறிமுகம்

தமிழைத் தமிழ் எழுத்துருக்களின் உதவியால் கணிப்பொறியில் எழுத இயலும். இக்காலக் கட்டத்தில், கணிப்பொறியில் எழுதப்பட்ட தமிழ்ப்பனுவல்களில் வரும் எழுத்துப் பிழைகளையும் இலக்கணப்பிழைகளையும் கணிப்பொறியே கண்டுபிடித்துத் திருத்தம் செய்வது இன்றியமையாத தேவையாகின்றது. எழுத்து மற்றும் இலக்கணப் பிழை திருத்தித் தயாரிக்கப் பின்வரும் உட்கருவிகள் தேவையாகுகின்றன.

1. கணினித் தொடரியல் ஆய்வி
2. கணினி உருபனியல் ஆய்வி
3. கணினி ஒலியனியல்/எழுத்து ஆய்வி
4. இயந்திரம் படிக்கவியலும் அகராதி

தமிழ்ப்பல்கலைக் கழக மொழியியல் துறையில் இவ்வாய்வு மேற்கொள்ளப்பட்டுள்ளது. தமிழுக்கு எழுத்துப்பிழை திருத்தி ஓரளவுக்குச் செம்மையாக வேலை செய்யும் வகையில் உருவாக்கப்பட்டது. இலக்கணப் பிழைதிருத்தி முழுவதுமாக உருவாக்கப்பெறவில்லை. அதற்கான ஆய்வுகள் மேற்கொள்ளப்பட்டு முயற்சிகள் எடுக்கப்பட்டுள்ளன.

சொல்லாய்வியின் வழி ஆவணங்கள் (documents) உருவாக்கப் படுகின்றன. ஆவணம் ஒரு கடிதமாகவோ சிறுகுறிப்பாகவோ கட்டுரையாகவோ நூலாகவோ இருக்கலாம். சொல்லாய்வியில் ஆவணத்தை உருவாக்கப் பொதுவாகப் பின்வரும் மூன்று முறைகள் பயன்படுத்தப்படுகின்றன:

1. கணிப்பொறியின் விசைப்பலகை (keyboard) வழி உள்ளீடு செய்தல்
2. பேச்சுணரி ஒழுங்கு முறையைப் (Voice Recognition System (VRS)) பயன்படுத்தி உரையை உள்ளீடு செய்தல்: இவ்வாறு செய்யும் பொழுது ஒலிவடிவில் பேசப்படும் செய்திகள் உரைவடிவிற்கு மாற்றப்படுகின்றன. இதற்குப் பேச்சிலிருந்து உரை (Speech to Text (STT)) என்ற தொழில் நுட்பம் உதவுகிறது.
3. தட்டச்சு செய்தல் அல்லது அச்சடிக்கப்பட்ட உரையை ஒளிவழி எழுத்துணர்தல் (Optical Character Recognition (ORC)) என்ற தொழில்நுட்பத்தின் மூலம் கணிப்பொறிக்கு உள்ளீடு செய்து சொல்லாய்விற்கு ஆவணமாக மாற்றுவது.

உரையை உள்ளீடு செய்ய மேற்கண்ட மூன்று முறைகளுள் ஏதேனும் ஒன்றைப் பின்பற்றலாம். பின்னர் சொல்லாய்வியின் மூலம் அவைகளைக் கையாளலாம். உரைக்குள்

எழுத்துப் பிழைகளும் இலக்கணப் பிழைகளும் இருப்பதற்கான வாய்ப்புகள் உள்ளன. ஆங்கில உரைகளில் உள்ள எழுத்துப் பிழைகளைக் கணிப்பொறிதானாகவே புரிந்துக்கொண்டு மாற்றுவதற்குச் சொல்லாய்வியில் வசதிகள் உள்ளன. இதற்கு எழுத்துப் பிழைகளைக் கண்டறிந்து திருத்துதல் என்று பெயர். இத்தகைய பிழைத்திருத்தி தமிழ் சொல்லாய்விகளில் உருவாக்க முயற்சிகள் எடுக்கப்பட்டு வருகின்றன. எழுத்துப் பிழைகள் நிகழ்வதற்கான வாய்ப்புகள், எழுத்துப் பிழைகளின் வகைகள், பிழைகளைக் கண்டறிந்து திருத்துவதற்கான செயல்பாட்டு வழிமுறைகள் போன்றவையெல்லாம் விரிவாக ஆராய்ந்து செயல்படவேண்டி உள்ளது.

பிழைகளை எழுத்துப்பிழைகள், இலக்கணப்பிழைகள் என்று வரையறுப்பது கவனக்குவிப்பு செய்வதற்கு எளிமையாக அமைந்தாலும் இரண்டையும் வெவேறான செயல்பாடுகள் என்று கூறுவது கடினம். எழுத்துப்பிழைகள் இலக்கணப்பிழைகளுக்கு வழிவகுக்கலாம் அல்லது இலக்கணப்பிழைகள் எழுத்துப்பிழைகளாகத் தோன்றலாம். குறிப்பாகத் தமிழ் போன்ற திரிபு ஒட்டுகள் மிகுந்த மொழியில் எழுத்துப்பிழை இலக்கணப்பிழைக்கும் இலக்கணப்பிழை எழுத்துப்பிழைக்கும் வழிவகுக்கும். சந்திப் பிழைகளை அல்லது புணர்ச்சிப் பிழைகளை எழுத்துப்பிழைகளாகவும் இலக்கணப்பிழைகளாகவும் கருதலாம்.

7.2. எழுத்துப்பிழை திருத்தி

எழுத்துப் பிழைகளைக் கண்டுபிடிப்பதும் திருத்துவதும் தற்கால சொல்லாய்விகளின் ஒருங்கிணைந்த பகுதியாக அமைகின்றது. எழுத்துப்பிழை திருத்தம் எழுத்துப்பிழை நிகழ்வதற்கான வாய்ப்புகளைக் கண்டறிதல், எழுத்துப்பிழைகளக் கண்டறிதல், பிழை நீக்கும் வழிமுறைகளைக் கண்டறிதல், பிழை நீக்குவதில் உள்ள சிக்கல்களை அறிதல் என்பன போன்ற செயல்பாடுகளை உட்படுத்தும்.

7.2.1. எழுத்துபிழைகள் நிகழ்வதற்கான வாய்ப்புகள்

உரைச் செய்திகளைக் கீழ்கண்ட ஏதேனும் ஒருங்குமுறைகளில் உள்ளீடு செய்யப்படும்போது எழுத்துப்பிழைகள் நிகழ்வதற்கு வாய்ப்புகள் உள்ளன.

- விசைப்பலகை வழி உள்ளீடு செய்யும்பொழுது தட்டச்சுப் பிழைகள் (typographical errors) ஏற்படுகின்றன.
- பேச்சுணரி ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்றபொழுது சொல்லின் உச்சரிப்பு மாற்றங்களினால் பிழைகள் ஏற்படுகின்றன. இத்தகைய பிழைகளுக்கு உச்சரிப்பு வேறுபாடு (pronunciation variation) என்று பெயர்.
- ஒளி எழுத்துணரி (ORC) வழி உள்ளீடு செய்கின்றபொழுது கருவிகளின் குறைபாடுகளினாலோ அல்லது எழுத்துக்களைச் சரியாகக் கருவிகள் படித்தறிய இயலாததாலோ பிழைகள் தோன்றுகின்றன. எடுத்துக்காட்டாக ஆங்கிலத்தில் l என்ற

எழுத்திற்கும் 1 என்ற எழுத்திற்கும் வேறுபாடு தெரியாததால் பிழை ஏற்படுகின்றன. பின்வரும் அட்டவணையில் சில சரியான எழுத்துக்களும் அவற்றிற்கிணையான தவறாகப் புரிந்துகொள்ளப்படும் எழுத்துக்களும் பட்டியலிடப்பட்டுள்ளன:

சரியான எழுத்து	**தவறாகப் புரிந்துகொள்ளப்பட்ட எழுத்து**
l	1
D	O
ri	n
0	θ

மேற்கண்டவாறு மூன்றுவகையான உள்ளீட்டு முறைகளிலும் அதற்கான கருவிகளைப் பயன்படுத்தும் பொழுது பிழைகள் தோன்றுகின்றன. பேச்சறிதல் ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்றபொழுது ஒரே வார்த்தையை அதன் இலக்கணக் கூறு மற்றும் தன்மைகளுக்கு ஏற்ப வெவ்வேறு விதமாக உச்சரிக்க வேண்டியுள்ளது. எடுத்துக்காட்டாக ஆங்கிலத்தில் object என்ற சொல் பெயர்ச் சொல்லாகவும் வினைச் சொல்லாகவும் பயன்படுகின்றது. பெயர்ச்சொல்லாக பயன்படும்போது 'ஆப்ஜிக்ட்' என்றும் வினைச் சொல்லாகப் பயன்படுத்தும்போது 'அப்ஜெக்ட்' என்றும் உச்சரிப்பு செய்யவேண்டியுள்ளது. அதைப்போல் the என்ற சொல்லைச் சூழலுக்கு ஏற்ப [ðə] என்றும் என்றும் [ði] மாறுப்பட்ட இருவகைகளில் உச்சரிக்க வேண்டியுள்ளது. இவ்வாறாக, ஒரே சொல்லிற்கான உச்சரிப்பு மாற்றங்களுக்குத் திட்டவட்டமாக நன்கு வரையறுக்கப்பட்ட விதிமுறைகள் உள்ளன. அவ்விதி முறைகளைக் கணிப்பொறிக்கு ஏற்ற வண்ணம் வடிவமைப்புகளாக வடிவமைத்து இத்தகைய பிழைகள் நீக்கப்படுகின்றன.

7.2.2 எழுத்துப் பிழைகளைக் கண்டறிதல்-சரிசெய்தல் சிக்கல்களின் வகைகள்

குகிச் (Kukich, 1992) தமது மதிப்பீட்டாய்வு கட்டுரையில் எழுத்துப்பிழை திருத்தத்தைப் பரந்த அளவில் மூன்றாகப் பிரிகின்றார். இதன் படி எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் தொடர்பான சிக்கல்கள் பின்வரும் செயல்பாடுகளை உள்ளடக்கும்:

- மொழியில் இல்லாத சொற்களைக் கண்டறிதல் (Non-Word Error Detection)
- எழுத்துப்பிழை உள்ள செய்திகளை மட்டும் தனிமைப்படுத்தி ஆராய்தல் (Isolated-word error correction)
- எழுத்துப்பிழை உள்ள சொல்லின் முன்பின் உள்ள அல்லது அச்சொல் இடம்பெற்ற தொடரின் பொருளையமாகக் கொண்டு பிழைநீக்கி சரிசெய்தல் (context-dependent error detection and correction)

குகிச் (Kukich 1992) மனிதர்களின் தட்டச்சுப் பிழைகளை இரு பிரிவுகளாகப் பிரிக்கிறார்:

1. தட்டச்சுப் பிழைகள் (typographic errors)
2. புரிதல் பிழைகள் (cognitive errors) சொல்லின் எழுத்து வடிவம் உரையாளருக்குத் தெரியாததால் ஏற்படும் பிழைகள்

டமரு (Dameraru, 1964) தவறாக எழுதப்பட்ட சொற்களில் 80% தனிநிலைத் தவறால் ஏற்பட்ட பிழைகளாகக் காண்கிறார். இதன் படி தட்டச்சுப் பிழைகளைப் பின்வரும் நான்கு வகைகளாகப் பிரிக்கலாம்:

1. விடுபடு பிழை (Deletion Error): இப்பிழையில் ஒரு எழுத்து விடுபட்டிருக்கும். எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் *'ட்'* என்ற எழுத்தில்லாமல் *'சடம்'* என இருப்பது விடுபடு பிழையாகும்.
2. செருகல் பிழை (Insertion Error): இப்பிழையில் ஒரு எழுத்து கூடுதலாக இருக்கும். எடுத்துக்காட்டாக, *சட்டடம்* என்ற பிழையான சொல்லில் *'ட'* என்ற எழுத்துக் கூடுதலாக இருப்பது செருகல் பிழையாகும்.
3. பதிலீட்டுப் பிழை (Substitution Error): இப்பிழையில் ஒரு எழுத்திற்குப் பதிலாக வேறொரு எழுத்தை மாற்றி அமைத்து விடும். எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் முதல் எழுத்து *'ச'* தவறாக அடிக்கப்படும்போது அத்தவறான எழுத்து 'ச' ஆகவோ 'க' ஆகவோ 'ப' ஆகவோ இருக்கலாம். இம்மாதிரியான பிழைகளைப் பதிலீட்டுப் பிழைகள் எனலாம்.
4. இடமாற்றுப் பிழை (Transposition Error): இப்பிழையில் அடுத்தடுத்த இரண்டு எழுத்துக்கள் இடம் மாற்றி இருக்கும். எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் பிழையாக *'ட'* என்ற எழுத்தும் *'ட்'* என்ற எழுத்தும் இடம் மாறி *சடட்ம்* என்று மாற்றி அடிக்கப்படுவது இடமாற்றுப் பிழையாகும்.

7.2.3 மொழியில் இல்லாத சொற்களைக் கண்டறிந்து மாற்றுதலுக்கான செயல் திட்டம்

ஒரு மொழியில் உள்ள சொல்லினைக் கண்டறிந்து சரிசெய்வதற்கு அகராதி பயன்படுத்தப்படுகிறது. கொடுக்கப்பட்ட சொல் அகராதியில் இல்லை எனில் அச்சொல்லை மொழியில் இல்லாத சொல் எனக் கருதலாம். இதற்குத் தகுந்த கணிப்பொறிக்கான அகராதி சிறியதாக இருத்தல் நலம். மிகப் பெரிய அகராதியாக இருந்தால் அதிகமாகப் புழக்கத்தில் உள்ள எழுத்துப்பிழைச் சொற்கள் கூட அரிதாகப் பயன்படுத்தும் சரியான சொல்லாக இருக்கும்.

7.2.4. பிழை நீக்கும் வழிமுறைகள்

உரைச் செய்திகளில் உள்ள எழுத்துப் பிழைகளைக் கண்டறிந்து மாற்றுவதற்கு பின்வரும் இரு நிகழ்தகவு மாதிரிகள் (probalistic models) பயன்படுத்தபடுகின்றன:

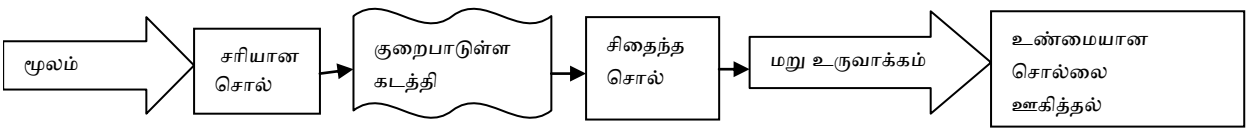
1. பெய்சியன் விதிகள் (Bayesian Rules)
2. ஓசை ஊடக மாதிரி (Noisy Channel Model)

பெய்சியன் விதிமுறைகளைப் பின்பற்றும் பொழுது எழுத்துப் பிழைகளைக் கண்டறிவதற்கான வாய்ப்புகளை ஆராய்வதற்குக் கணித முறையில் அமைந்த நிகழ்தகவுக் கோட்பாடு (Probability Theory) பயன்படுத்தப்படுகிறது. பெரும்பாலும் இம்முறைகள் தட்டச்சு செய்த உரைகளுக்குப் பயன்படுத்தப்படுகின்றன. இம்முறையை ஒளிவழி எழுத்துணரி (OCR) முறையிலும் பேச்சுப் புரிந்துகொள்ளல் ஒழுங்குமுறைகளிலும் (Voice Recognition System) உள்ளீடு செய்யப்படும் உரைகளுக்கும் பயன்படுத்தலாம்.

இரண்டாவது ஓசை ஊடக மாதிரி (Noisy Channel Model) முறையில் உள்ளீடு செய்யப்படுகின்ற உரைகள் எவ்வாறு மாற்றமடைகின்றன, மாற்றமடைவதற்கான காரணங்கள் யாவை போன்றவைகள் ஆராயப்பட்டு ஒரு மாதிரி வடிவம் (model) தயாரிக்கப்படுகின்றது. பின்னர் இம்மாதிரி வடிவத்திற்கான எதிர்மறைச் செயல்பாடுகள் (reverse process) ஒருமாதிரிப் படிவமாக உருவாக்கப்பட்டுப் பிழைகள் நீக்கப்படுகின்றன. எடுத்துக்காட்டாக, தொலைபேசியில் பேசுகின்றபொழுதும் ஒலிவாங்கி வழி பேசுகின்றபொழுதும் ஒருவர் பேசும் குரல் அமைப்புகள் மாறுபடுகின்றன. இத்தகைய கருவிகள் மூலம் கிடைக்கப்பெறுகின்ற முடிவுகள் ஒலி அமைப்பைப் பொறுத்துள்ளது; உள்ளபடியாகவே அமையாது. இந்நிலையில் அம்மாறுபாடுகள் நிகழ்வதற்கான காரணங்களை ஆராய்ந்து அதற்கு எதிர்மறையான திட்டங்களை வடிவமைக்கின்ற பொழுது இவைகள் மூலம் நிகழும் பிழைகளை நீக்கலாம்.

7.2.4.1 பெசியன் முறை (Bayesian Medhod)

ஒரு உரையிலுள்ள எழுத்துப்பிழைகளைக் கண்டறிந்து திருத்துவதற்குப் பல செயல் திட்ட முறைகள் வடிவமைக்கப்படுகின்றன. அவற்றில் பெசியன் முறை நிகழ்தகவு அடிப்படையில் (Probability Thories) அமைந்ததாகும். இம்முறையின் அடிப்படைத் தத்துவம் ஜுராஸ்கி மற்றும் மார்டினிலிருந்து எடுத்தாளப்பட்ட பின்வரும் படத்தில் விளக்கப்பட்டுள்ளது (Jurafsky and Martin 2000: 147).



உரைக்கான சொற்கள் தகுந்த உள்ளீட்டுக் கருவிகள் மூலம் உள்ளீடு செய்யப்படுகின்றன. உள்ளீட்டுக் கருவி கணிப்பொறியின் விசைப்பலகையாகவோ ஒலியுணரி ஒழுங்குமுறையின்

ஒலிவாங்கியாகவோ இருக்கலாம். தட்டச்சு செய்யும்பொழுது எழுத்துக்களின் சிதைவோ ஒலியுணரி ஒழுங்குமுறையில் ஒலிச் செய்தியின் சிதைவோ ஏற்பட வாய்ப்புகள் உள்ளன. மூலக்கருவிகள் மூலம் உருவாக்கப்படும் சொல் உரிய கடத்திகள் (channels) மூலம் கடத்தப்படுகிறது. கடத்திகளில் உள்ள குறைபாடுகள் காரணமாகவும் கடத்தும் செய்திகளில் சிதைவு ஏற்படலாம். இவ்வாறாகச் சிதைந்த சொல் மறு உருவாக்கம் (decode) செய்யப்படுகிறது. இச்சொல் மூலக்கருவி மூலம் உருவாக்கப்பட்ட சொல்லுக்கு வாய்ப்புள்ள சொல்லாகவே அமையும். இவ்வாறு சிதைந்த சொல்லுக்கு வாய்ப்புள்ள சொற்களின் பட்டியல் உருவாக்கப்படும். இப்பட்டியலிலிருந்து சிதைந்த சொற்கூறு திருத்தம் செய்யப்பெற்றுப் பிழை நீக்கிய சரியான சொல் அடையாளம் காணப்படுகிறது.

பேசியன் முறையில் எழுத்துத் திருத்திக்கான வழிமுறைகள் இருபெரும் கட்டங்களாகச் செய்யப்படுகின்றன:

1. பிழை நீங்கிச் சரியான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியல் தயாரித்தல்

2. இச்சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லுக்கும் மதிப்பெண் அளித்து அதிக மதிப்பெண் பெறும் சொல்லைப் பிழைத் திருத்தம் செய்யப்பட்ட சொல்லாகக் கருத வழிவகை செய்தல்.

இம்முறையில் பிழைத்திருத்தம் என்பது சொல்லில் ஏதேனும் ஒரு எழுத்து மட்டும் பிழையாக இருக்கலாம் என்ற வாதத்தின் அடிப்படையில் செயல்முறைகள் மேற்கொள்ளப்படுகின்றன. எனவே பின்வரும் வகையிலான எழுத்து பிழைகள் மட்டும் இம்முறையில் எடுத்துக் கொள்ளப்படுகின்றன:

- சொல்லின் ஒரு எழுத்து விடுபட்டு இருத்தல்
- சொல்லில் ஒரு எழுத்து அதிகமாக இருத்தல்
- சொல்லில் ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து தவறாக இருத்தல்
- சொல்லில் அடுத்தடுத்த இரு எழுத்துக்கள் இடம் மாறி இருத்தல்

பேசியன் முறையில் எழுத்துத் திருத்தி உருவாக்குவதற்கு முதலில் பெருந்தரவு (Corpus) ஒன்றினைக் கணிப்பொறியில் உருவாக்கிக்கொள்ள வேண்டும். இந்தப் பெருந்தரவில் மொத்தம் ‘N’ சொற்கள் இருப்பதாகக் கொள்வோம். பிழையானச் சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியல் ஒன்றை உருவாக்க வேண்டும். இப்பட்டியலை உருவாக்கும் ஒரு சொல்லில் ஒரு எழுத்தினால் மட்டும் பிழை நிகழ்வதாக எடுத்துக்கொள்ளப்படுகிறது. எனவே சொல்லில் (அ) ஒரு எழுத்து விடுபட்டு இருத்தல், (ஆ) ஒரு எழுத்து கூடுதலாக இருத்தல், (இ) ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருத்தல், (ஈ) அடுத்தடுத்த இரு எழுத்துக்கள் தவறாக இடம் மாறியிருத்தல் ஆகிய நான்கு மட்டும் கருத்தில் கொள்ளப்படுகின்றன. மேற்கண்ட முறைகளின்படி பிழையான

சொல்லிற்கு வாய்ப்புள்ள சொற்பட்டியலில் 'V' சொற்கள் இருப்பதாகக் கொள்வோம். இப்பட்டியலில் 'c' என்ற சொல்லாக இருப்பதற்கான முன்நிகழ்வு எண்ணிக்கை பின்வரும் சூத்திரத்தின் மூலம் கணக்கிடப்படுகிறது (Jurafsky and Martin 2000: 151):

$$P(c) = \frac{C(c)+0.5}{N+0.5V}$$

(V = வாய்ப்புள்ள சொல், C = விரிதரவு 'Corpus')

மேற்கண்ட சூத்திரம் ஒரு சொல் மட்டும் பெருந்தரவில் இடம்பெற்றுள்ள முன்நிகழ்வு (Prior) மதிப்பைத் தருகிறது. இரண்டாவது கட்டநிலையில் தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்களுக்கு மதிப்புகள் கண்டறியப்படுகின்றன. இதற்கு மாதிரி உரைகள் ஒரு தட்டச்சர் அல்லது பல தட்டச்சர்களைக் கொண்டு உள்ளீடு செய்யப்படுகின்றன. கெர்னிகன் மற்றும் பிறர் (Kernighan et al 1990) ஒவ்வொரு வகைத் தனிநிலைப் பிழைக்கும் நான்கு குழப்ப அணிகளைப் (confusion matrices) பயன்படுத்தினர்.

1. Del (x, y) என்பது xy என்ற இரு எழுத்துகளுக்குப் பதிலாக x மட்டும் எத்தனை முறை தவாற அச்சடிக்கப்பட்டுள்ளது என்பதைக் குறிக்கிறது.
2. Ins (x, y) என்பது x என்ற எழுத்துக்களுக்குப் பதிலாக xy என்று தட்டச்சு செய்துள்ள எண்ணிக்கையைக் குறிக்கிறது.
3. Sub (x, y) என்பது x என்ற எழுத்திற்குப் பதிலாக எத்தனை முறை y என்ற எழுத்து தட்டச்சு செய்யப்பட்டுள்ளது என்பதைக் குறிக்கிறது.
4. Trans (x, y) என்பது xy என்ற எழுத்துக்களுக்குப் பதிலாக yx என இடமாறி அச்சடிக்கப்பட்டுள்ளது என்பதைக் குறிக்கிறது.

தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியலில் இருந்து ஏதாவது ஒரு சொல் 'c' என்பதை எடுத்துக்கொள்வோம். 'c_p' என்பது 'c' என்ற சொல்லில் P- இடத்தில் உள்ள எழுத்தைக் குறிப்பதாகக் கொள்வோம். இந்நிலையி 'c' என்ற சொல் வாய்ப்பான சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது (Jurafsky and Martin 2000: 152):

1. ஒரு எழுத்து விடப்பட்டிருக்கும் நிலையில்:

$$P(t/c) = \frac{del[c_p - 1, c_p]}{count[c_{p-1} c_p]}$$

2. ஒரு எழுத்து கூடுதலாக இருக்கும் நிலையில்

$$P(t/c) = \frac{ins\ [c_{p-1}, t_p]}{count\ [c_{p-1}]}$$

3. ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருக்கும் நிலையில்

$$P(t/c) = \frac{trans\ [t_p, c_p]}{count\ [c_p]}$$

4. இரு எழுத்துக்கள் இடம் மாறி இருக்கும் நிலையில்

$$P(t/c) = \frac{trans[c_p, c_{p+1}]}{count\ [c_p, c_{p+1}]}$$

தவறான சொல் வாய்ப்புள்ள சொற்பட்டியலில் உள்ள 'c' என்ற சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது (Jurafsky and Martin 2000: 150):

$$W = P(t/c) * P(c)$$

இந்த சூத்திரத்தில் P(c) என்பது c என்ற சொல் பெருந்தரவில் உள்ள முன் நிகழ்வு மதிப்பைத் தருகிறது. P(t/c) என்பது தவறாகத் தட்டச்சு செய்யப்பட்ட சொல் 'c' என்ற சொல்லாக இருப்பதற்கு எவ்வளவு வாய்ப்புகள் உள்ளன என்ற மதிப்பைத் தருகின்றது. இவ்விரு எண்களின் பெருக்கல் தொகை, அதாவது P(t/c)*P(c) என்பது இச்சொல்லிற்கான மதிப்பெண்ணைத் தருகிறது.

மேற்கண்டவாறு தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லிற்கும் மேற்கண்ட மதிப்பினைக் கண்டறிய வேண்டும். எந்தச் சொல் அதிக மதிப்பெண் பெற்றிருக்கிறதோ அந்தச் சொல் மிகச் சரியான சொல் என கருதப்படுகிறது. அதிக மதிப்பெண் பெற்றிருக்கும் சொல் பின்வரும் சூத்திரத்தின் மூலம் கூறப்பட்டுள்ளது (Jurafsky and Martin 2000: 150).

$$W = \underset{c \in V}{\operatorname{argmax}} \frac{P(t/c)P(c)}{P(t)}$$

மேற்கண்டவாறு அதிக மதிப்பெண் பெற்ற சொல் இடம் பெற்ற வாக்கியத்தின் பொருளுடன் ஒப்பிடுப் பார்க்கப்படும். அது வார்த்தையின் பொருளுக்கேற்ற சொல் இல்லையெனில் அடுத்த மதிப்பெண் பெற்ற சொல் பொருத்திப்பார்க்கப்படும். இவ்வாறாகத் தவறாகத் தட்டச்சு செய்யப்பட்டுள்ள சொற்களைச் சரியான சொற்களாக மாற்றச் செயல் திட்டங்கள் உள்ளன.

இவ்வாறாக உரைகள் எழுதும்போது பிழைகளைக் கண்டறிந்து திருத்துவதற்குக் கணிப்பொறியில் பல செயல்திட்டங்கள் உருவாக்கப்படுகின்றன. அதில் மிகவும் பிரபலமானது மேற்கண்ட பேசியன் முறையாகும். இதுபோல் குறைந்த திருத்த தூரம் (minimum edit distance) என்ற மற்றொரு முறையும் பயன்படுகிறது.

7.2.4.2. குறைந்த திருத்து தூரம்

முந்தைய பகுதியில் பேசியன் வழிமுறைவரைவு எவ்வாறு திருத்தங்களைச் செய்ய முடிந்தது என்று பார்த்தோம். இருப்பினும் கெர்னிகன் மற்றும் பிறர் (Kernighan et al 1990) ஒரு சொல் ஒரு எழுத்துப்பிழை மட்டும் கொண்டிருக்கும் என்ற எளிய ஊகத்தைச் சார்ந்திருந்தனர். நாம் பன்முகப் பிழைகளைக் கையாளுவதற்குக் கூடுதல் ஆற்றலுள்ள வழிமுறை வரைவை விருப்பலாம். அம்மாதிரியான வழிமுறைவரைவைக் கோர்வை தூரத்தின் சிக்கலுக்குரிய பொதுவான தீர்வாகக் கருதலாம். 'கோர்வை தூரம்' என்பது எந்த அளவுக்கு இரண்டு கோர்வைகள் ஒன்றுபோல் இருக்கின்றன என்பதன் ஒரு அணியாகும். பேசியன் முறையைப் பிழைதிருத்தச் சிக்கலுக்கு அம்மாதிரியான வழிமுறைவரைவைப் பயன்படுத்தும் வழியாகப் பார்க்க இயலும் ; நாம் தரப்பட்ட பிழைக்கு மிக உயர்ந்த நிகழ்தகவு கொண்டது என்ற அர்த்தத்தில் தவறுக்கு மிக அண்மையில் உள்ள சொல்லைத் தேர்வுச் சொல்லாக எடுக்கின்றோம்.

கோர்வை தூரத்தைக் கண்டுபிடிப்பதற்கு வழிமுறைவரைவுகளின் பிரபலமான வகுப்புகளில் ஒன்று வாக்னர் மற்றும் ஃபிஷர் (Wagner and Fischer 1974) என்பவர்களால் குறைந்த திருத்து தூரம் (minimum edit distance) என்று பெயரிடப்பட்ட (ஆனால் தனித்தனியாகப் பலரால் கண்டுபிடிக்கப்பட) ஒரு பதிப்பாகும். இரு கோர்வைகளுக்கு இடையே உள்ள குறைந்த திருத்து தூரம் என்பது ஒரு கோர்வையை மற்றொரு கோர்வையாக மாற்றத் தேவைப்படும் திருத்து நடவடிக்கைகளின்/செயன்மைகளின் (செருகல், நீக்குதல், பதிலீடுசெய்தல்) குறைந்த எண்ணிக்கையாகும். எடுத்துக்காட்டாக intension மற்றும் extension என்ற சொற்களுக்கு இடையே உள்ள இடைவெளி ஐந்து செயன்மைகளாகும்/நடவைக்கைகளாகும்; இதை மூன்று வழிகளில் உருப்படுத்தம் செய்யவியலும் ; கீழே காட்டியுள்ளது போன்ற சுவடு/தடம் (trace), வரிசையாக்கம் (alignment) அல்லது செயன்மைப் பட்டியல் (operation list) என்ற மூன்று வழிகள் உள்ளன (Jurafsky and Martin 2000: 154).

சுவடு

I n t e n t i o n
/ / / / | | | |
e x e c u t i o n

வரிசையாக்கம்

I n t e n ε t i o n
ε e x e c u t i o n

நடவடிக்கைப் பட்டியல்

I n t e n t i o n

I-ஐ நீக்கவும்

n t e n t i o n

n-ஐ e-ஆல் பதிலீடு செய்யவும்.

e t e n t i o n

t-ஐ x-ஆல் பதிலீடு செய்யவும்.

e x e n t i o n

u-ஐ செருகவும்

e x e n u t i o n

n-ஐ c-ஆல் பதிலீடு செய்யவும்

e x e c u t i o n

நாம் ஒவ்வொரு நடவடிக்கைக்கும் ஒரு குறிப்பிட்ட மதிப்பையோ (cost) சுமையையோ (weight) ஒதுக்கலாம். இரு வரிசைகளுக்கு இடையிலான லெவென்ஷ்டைன் தூரம் தான் மிக எளிய அளவிடும் காரணி; இதில் ஒவ்வொரு நடவடிக்கைக்கும் விலை 1 (Levenshtein 1966) ஆகும். இதன்படி intension என்பதற்கும் execution என்பதற்கும் இடையில் உள்ள தூரம் 5 ஆகும். லெவென்ஷ்டைன் தன்னுடைய அளவுமுறைக்கு மற்றொரு மாற்று பதிப்பை முன்மொழிகின்றார்; இதில் ஒவ்வொரு செருக்கலுக்கும் அல்லது நீக்கலுக்கும் மதிப்பு ஒன்று ஆகும்; பதிலீடு அனுமதிக்கப்படமாட்டாது (பதிலீட்டை ஒரு செருகலாலும் ஒரு நீக்கலாலும் உருப்படுத்தம் செய்ய இயலும் ஆகையால் ஒவ்வொரு பதிலீட்டுக்கும் மதிப்பு 2 தரப்படுகின்றது). இந்தப் பதிப்பின் பயன்பாட்டின் படி intension என்பதற்கும் execution என்பதற்கும் இடையிலுள்ள தூரம் 8

ஆகும். நாம் கூடுதல் கலவைத்தன்மையான செயல்பாடுகளால் சுமையேற்றலாம்; எடுத்துக்காட்டாக ஒவ்வொரு நடவடிக்கைக்கும் முன்னர் கூறிய குழப்ப அணிகளைப் பயன்படுத்தி ஒரு நிகழ்தகவை தருவது. இந்நேர்வில் இரண்டு கோர்வைகளுக்கு இடையிலுள்ள குறைந்த திருத்து தூரத்தைப் பற்றி பேசுவதற்குப் பதிலாக நாம் ஒரு கோர்வைக்கும் மற்றொரு கோர்வைக்கும் அதிக நிகழ்தகவு வரிசையாக்கம் பற்றி பேசுவோம். இதை நாம் செய்தால் ஒவ்வொரு மாற்றத்தின் நிகழ்தகவுகளைப் பெருக்கும் பெரிதாக்கப்பட்ட குறைந்த திருத்து தூர வழிமுறைவரைவை ஒரு தேர்வுப் பிழை தரப்படுகையில் பல் தவறு தட்டச்சின் பேசியன் சாத்தியத்தியத்தை மதிப்பிட பயன்படுத்த இயலும்.

ஆற்றலுள்ள வழிமுறை அமைப்பால் (dynamic programming) குறைந்த திருத்து தூரம் கணக்கிடப்படுகின்றது. ஆற்றலுள்ள வழிமுறை அமைப்பு வழியமைப்பு வரைவுகளின் ஒரு வகுப்பின் பெயராகும்; இது முதலில் பெல்மான் (Bellman 1957) என்பவரால் அறிமுகப்படுத்தப்பட்டது; இது துணைச் சிக்கல்களின் திர்வுகளை ஒன்றிணைத்து சிக்கல்களைத் தீர்க்க அட்டவணை-இயக்க முறையைப் (table-driven method) பயன்படுத்துகின்றது. இந்த வழியமைப்பு வரைவுகளின் வகுப்பு பேச்சு மற்றும் மொழி ஆய்வுகளில் கூடுதல் பொதுவாகப் பயன்படுத்தப்படுகின்ற வழியமைப்பு வரைவுகளை உட்படுத்தும்; எழுத்துப் பிழை திருத்ததிற்கான குறைந்த திருத்து தூர வழியமைப்பு வரைவு, பேச்சு புரிந்துகொள்வதிலும் இயந்திர மொழிபெயர்ப்பிலும் பயன்படுத்தப்படும் வெட்டர்பி வழியமைப்பு வரைவு (Veterbi algorithm) மற்றும் பகுத்துக்குறிக்கப் பயன்பத்தப்படும் CYK மற்றும் Early வழியமைப்பு வரைவு என்பன இவையாகும்.

7.3. இலக்கணப்பிழைத் திருத்தி

இலக்கணத் திருத்தி ஒரு வாக்கியத்தின் தொடரியல் செம்மையை நிர்ணயிக்கும் அல்லது தீர்மானிக்கு இலக்கணத் திருத்தம் சொல் ஆய்விகளிலும் (word processor) தொகுப்பிகளிலும் (compilers) பெரும்பாலும் பயன்படுத்தப்படுகின்றது. தொகுப்பி போன்ற பயன்பாடுகளில் இலக்கணத் திருத்தம் எளிமையானதாகும். எனென்றால் சொற்றொகை அவற்றில் எல்லைக்குப்பட்டதாகும்; ஆனால் இயற்கை மொழிகளில் இது சவாலுக்குரியதாகும்; ஏனென்றால் அதன் சொற்றொகை எல்லைக்கு உட்பட்டதல்ல.

மொழிகளில் இலக்கணத் திருத்தத்திற்கு மூன்று நெறிமுறைகள் பயன்படுத்தப்படுகின்றன; அவையாவன:

(1) தொடரியல் அடிப்படையிலான இலக்கணத்திருத்தம் (syntax-based checking)

(2) புள்ளியியல் அடிப்படையிலான இலக்கணத்திருத்தம் (Statistics-based checking)

(3) விதி அடிப்படையிலான இலக்கணத்திருத்தம் (Rule-based checking)

7.3.1 தொடரியல் அடிப்படையிலான திருத்தம் (Syntax based checking)

தொடரியல் அடிப்படையிலான இலக்கணத் திருத்தத்தில் ஒவ்வொரு வாக்கியமும் அவற்றின் இலக்கணச் செம்மைக்காக முழுமையாகப் பகுத்தாயப்படும் (parsed). தொடரியல் பகுத்துக்குறித்தல் பிழையானால் பனுவல்/உரை தவறானதாகக் கருத்தப்படும். ஜென்சன் (Jensan 1993) என்பவரால் உருவாக்கப்பட்ட அணுகுமுறையயில் ஒரு உரை முழுவதும் பகுத்துக்குறிக்கப்படும். அதாவது வாக்கியங்கள் ஆயப்பட்டு ஒவ்வொரு வாக்கியத்திற்கும் ஒரு கிளை அமைப்பு தரப்படும். பகுத்துக்குறித்தல் வெற்றிகரமாக நடைபெறாவிட்டால் உரையானது தவறானதாகக் கருதப்படும்.

தொடரியியல் அடிப்படையிலான அணுகுமுறையின் சிறப்பு என்னவென்றால் இலக்கணம் முழுமையாக இருந்தால் இலக்கணத் திருத்தலும் முழுமையாக அமையும். அதாவது இலக்கணத் திருத்தி தவறு எவ்வளவு தெளிவற்று இருந்தாலும் எந்தத் தவறான வாக்கியத்தையும் கண்டுபிடிக்கும். துரதிஷ்ட வசமாக இலக்கணத் திருத்தி தவறான வாக்கியங்களை மட்டும் தெரிந்துகொள்ளும்; அது பயன்படுத்துவோருக்குச் சரியாக என்ன சிக்கல் என்பதைக் கூற இயலாது. இதற்காகக் கூடுதல் விதிகள் தேவைப்படும்; அதுவும் தவறான வாக்கியங்களைப் பகுத்துக்குறிக்கும். ஒரு வாக்கியத்தை அம்மாதிரியான கூடுதல் விதியால் மட்டும் பகுத்துக்குறிக்க இயலுமென்றால் அவ்வாக்கியம் தவறானதாகும். இந்த நுட்பம் கட்டுப்பாடு நெகிழ்வு (Constraint Relaxation) எனப்படும்.

எனினும், தொடரியல் சார்ந்த அணுகுமுறையில் ஒரு முக்கிய பிரச்சனை உள்ளது: அதற்கு ஒருவர் சரிபார்க்க விரும்பும் அனைத்து வகையான பனுவல்களையும் உள்ளடக்கிய ஒரு முழுமையான இலக்கணம் தேவை. பல இலக்கண கோட்பாடுகள் இருந்தாலும் இன்னும் வலுவான பரந்த உள்ளடக்கப்பரப்புப் பாகுபடுத்தி (robust broad-coverage parser) பலரறிய எதுவும் கிடைக்கவில்லை. மேலும் பாகுபடுத்திகள் இயற்கையான பொருண்மை மயக்கத்தால் பாதிக்கப்படுகின்றன; இதன் காரணமாகச் சரியன வாக்கியங்களுக்கும் வழக்கமாக ஒன்றுக்கும் மேற்பட்ட விளைவுகள் திரும்பக்கிடைக்கின்றன.

7.3.2. புள்ளியியல் அடிப்படையிலான திருத்தம் (Statistical based checking)

புள்ளியியல் அடிப்படையிலான இலக்கணத்திருத்தத்தில் சொல்வகைப்பாடு அடையாள (POS tag) வரிசைகள் ஒரு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியிலிருந்து உருவாக்கப்படும்; இவ்வரிசைகளின் நிகழ்வெண் அதாவது நிகழ்தகவு/சாத்தியக்கூறு (probability) குறித்துக்கொள்ளப்படும். சொல்வகை அடையாளப்படுத்தப்பட்ட உரை/பனுவல் சில

நிகழ்வெண்ணைக் காட்டிலும் குறைவான நிகழ்வெண்ணில் சொல்வகைப்பாடு வரிசைகளைக் கொண்டிருந்தால் அப்பனுவல் பிழையானதாகக் கருத்தப்படும். புள்ளியியல் அடிப்படை அணுகுமுறை முக்கியமாக அடையாளப்படுத்தப்பட்ட பயிற்சி தரவுத்தொகுதியிலிருந்து (tagged training corpus) விதிகளைக் கற்றுக்கொள்ளும். விதி அடிப்படை அணுகுமுறையில் விதிகள் கையால் எழுத்தப்படவேண்டும் என்பது தவிர அணுகுமுறை புள்ளியியல் அணுகுமுறையுடன் ஒற்றுமை உடையதாகும்.

இருப்பினும் ஆங்கிலத்திற்கு மிகவும் பயன்படுத்தப்படுகின்ற இலக்கணத்திருத்திகளில் ஒன்றான மைகிரோசாஃப்ட் ஆபீஸ் சூட் (Microsoft Office Suit) இலக்கணத் திருத்தியும் எதிர்கருத்திற்கு (controversy) அப்பாற்பட்டதல்ல. தற்போது இலக்கணத் திருத்தி உருவாக்கம் மிகவும் எளிமையான செயல் அல்ல என்பதை நிரூபிக்கின்றது.

இது அட்வல் (Attwell1987) என்பவரால் உருவாக்கப்பட்ட அணுகுமுறையில் சொல்வகைப்பாட்டுக்கு அடையாளப்படுத்தப்பட்ட தரவுத் தொகுதி சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட கோர்வைகளின் பட்டியலை உருவாக்கப் பயன்படுத்தப்படும். சில கோர்வைகள் பொதுவானதாகும்.

எ.கா.

அடைகொளி அடை + பெயரடை+ பெயர் = The old man.

பிற எல்லாவற்றிலும் நேர்வதில்லை.

எ.கா.

அடைகொளி அடை + பெயரடை

தரவுத் தொகுதியில் அடிக்கடி நேரும் கோர்வைகளைப் பிற உரைகளில் சரியானதாகக் கருத இயலும்; அரிதாக நேரும் கோர்வைகள் தவறானதாகக் கருதப்படும்.

புள்ளியியல் அணுகுமுறையில் என்கிராம் பகுப்பாய்வை (n-gram analysis) பயன்படுத்தி ஒரு வாக்கியத்தின் சாத்தியக்கூறை அளக்கவியலும். எடுத்துக்காட்டாக, He is playing என்ற வாக்கியத்தின் பைகிராம் சத்தியக்கூறு பின்வருமாறு அமையும்.

P(“He is playing.”) = P(He|<start>)*P(is|he)*P(playing|is)*P(.|playing)

இம்மூன்று சொற்களில் ஏதேனும் ஒன்று பயிற்சித் தரவுத்தொகுதியில் இல்லாவிடில், வாக்கியத்தின் சாத்தியக்கூறு, பெருக்குத்தொகையின் காரணமாகப் பூஜியமாக மாறும். நாம் இப்புள்ளியியல் நெறிமுறையில் சொற்களைக் கருத்திற்கொண்டால் மொழியின் எல்லாச் சொற்களையும் கொண்டுள்ள பெரிய தரவுத்தொகுதி நமக்குத் தேவைப்படும்.

இச்சிக்கலைத் தீர்ப்பதற்காக நாம் தனிச்சொற்களுக்குப் பதிலாகச் சொல்வகைப்பாட்டு அடையாளங்களைப் (POS tags) பயன்படுத்த இயலும். வேறுபாடு என்னவென்றால், முன்னர்

நாம் எச்சொற்கள் ஒரு சொல்லுக்குப்பின் வரும் கூடுதல் சாத்தியக்கூறைக் கொண்டிருக்கின்றன என்பதைப் பரிசோதித்தோம்; தற்போது எச்சொல்வகை அடையாளங்கள் ஒரு சொல்வகைப்பாட்டு அடையாளத்திற்குப்பின் வரும் கூடுதல் சாத்தியக்கூறைக் கொண்டிருக்கின்றன என்று பரிசோதிப்போம். நாம் சொல்வகைப்பாடு அடையாளங்களைப் பயன்படுத்தினால், சொற்கள் மாறிகளாகும் (variables).

"He is playing." என்ற முந்தைய எடுத்துக்காட்டு வாக்கியத்தை எடுத்துக்கொள்ளவும். சொல்வகைப்பாடு அடையாளப்படுத்தலுக்குப் பிறகு அவ்வாக்கியம் He/pps is/bez playing/vbg./." என மாறும். இப்போது வாக்கியத்தின் சாத்தியக்கூறை கணக்கிட நாம் சொல்வகைப்பாட்டு வரிசையைப் பயன்படுத்த இயலும்.

P(pps bez vbg.) = P(pps|<start>)*P(bez|pps)*P(vbg|bez)*P(.|vbg)

நாம் கூறுகிற இலக்கண திருத்தி பின்வருமாறு செயலாற்றுகின்றது.

1. வாக்கியத்தின் ஒவ்வொரு சொல்லுக்கும் ஒரு அடையாளம் தரவும்.
2. அடையாள வரிசையின் (tag sequence) சாத்தியக்கூறை தீர்மானிக்க என்கிராம் பகுப்பாய்வை பயன்படுத்தவும் (இதில் என்=3 அதாவது டிரை கிராம்)
3. சாத்தியக்கூறு சில நிகழ்வெண்ணிற்கு மேல் இருந்தால் வாக்கியம் இலக்கண அடிப்படையில் சரியானது என்று கருதப்படும். நமது மாதிரியில் சாத்தியக்கூறு பூஜியத்திற்கு அதிகமாக இருந்தால் வாக்கியம் சரியானது என்று கருதப்படும். இரண்டோ அதற்கு மேற்பட்ட அடுத்துவருகிற அடையாளங்கள் பொருந்த இயலவில்லை என்றால் (அதாவது இயைபின்றி இருந்தால்) வரிசையின் சாத்தியக்கூறு பூஜியமாகும்.

முதலில் நமக்கு ஒரு சொல்வகை அடையாளப்படுத்தி (POS tagger) தேவை. அது தானியக்கமாகச் சொற்களை அடையாளப்படுத்தும்; அல்லது நாம் வாக்கியத்தின் சொற்களைக் கையால் அடையாளப்படுத்தவேண்டும். பின்னர் அடையாள வரிசையின் சாத்தியக்கூறைத் தீர்மானிக்க (இரண்டு முந்தைய அடையாளங்களைப் பார்க்கும்) டிரை கிராம் மாதிரியைப் பயன்படுத்தவும்; இறுதியாக அவ்வடையாள வரிசையின் சாத்தியக்கூறின் அடிப்படையில் இலக்கணச் செம்மையின் தீர்மானத்தைச் செய்யவும்/எடுக்கவும். எடுத்துக்காட்டாக பிரவுண் தரவுத்தொகுதி (Brown Corpus) மற்றும் பிரில்லின் சொல்வகைப்பாடு அடையாளப்படுத்தியை (Brill's tagger), "He saw the book on the table" என்ற வாக்கியத்தின் கணக்கீடுகள் பின்வருவனவாகும்:

He/pps saw/vbd the/at book/nn on/in the/at table/nn ./.

P(pps|None None) = 0.0635486169593

P(vbd|None pps) = 0.213047910296

P(at|pps vbd) = 0.166456494352

P(nn|vbd at) = 0.48308668061

P(in|at nn) = 0.362738953306

P(at|nn in) = 0.350967938144

P(nn| in at = 0.44004695623

P(.|at nn) = 0.0847696646819

அடையாள வரிசையின் சாத்தியக்கூறு = 5.16478478489e-09

இலக்கணத்திருத்தியின் முடிவு = இவ்வாக்கியம் சாத்திய அடிபடையில் சரியானதாகும்.

நாம் உடன்பாட்டு இயைபு இல்லாத “He have the book I want” என்ற வாக்கியத்தை ட்ரைகிராம் அடிப்படையில் ஆய்ந்தால் இலக்கணத் திருத்தி பின்வரும் கணக்கீடுகளைத் தரும்.

He/pps have/hv the/at book/nn I/ppss want/vb./.

P(pps|none none) = 0.0635486169593

P(hv|none pps) = 0.0

P(at|pps hv) = 0.0

P(nn|hv at) = 0.0491712707182

P(ppss|at nn) = 0.00493575681605

P(vb|nn ppss) = 0.293785310734

P(.|ppss vb) = 0.0361445783133

அடையாள வரிசையின் சாத்தியக்கூறு =0.0

இலக்கணத்திருத்தியின் முடிவு = தவறானது அல்லது கண்டுபிடிப்பது சாத்தியமில்லை

புள்ளியியல் சார்ந்த பாகுபடுத்திகள் அவற்றின் முடிவுகளை விளக்குவது கடினம் என்ற இடரைக் கொண்டிருக்கின்றது: ஒரு தவறான எச்சரிக்கைப் பிழை இருந்தால் பயன்பாட்டாளர் குறிப்பிட்ட பிழைச் செய்தி இல்லாத காரணத்தால் ஏன் அவரது உள்ளீடு தவறாகக் கருதப்படுகிறது என்று ஆச்சரியப்படுவர். உருவாக்குபவர் கூட ஒழுங்குமுறையின் தீர்ப்பைப் புரிந்துகொள்ள ஒழுங்குமுறைக்குப் பயிற்சி அளிக்கப்பட்ட தரவுத்தொகுதியை அணுக வேண்டிவரும். மற்றொரு சிக்கல் யாராவது ஒருவர் பொதுவல்லாத மற்றும் தவறான கட்டுகளைப் பொதுவல்லாத ஆனால் சரியான கட்டுகளிலிருந்து பிரிக்கும் தொடக்க நிலையை நிறுவவேண்டும். இந்தச் செயல் 0-வுக்கும் 100-க்கும் இடையில் ஏதாவது மதிப்பை நிறுவும் பயன்பாட்டாளருக்குத் தரப்படவேண்டும். தொடக்க நிலையின் கருத்து (நடை மற்றும் கட்டப்பட்ட இறுதி நேர்வுகள்

தவிர) வாக்கியங்கள் வழக்கமாகச் சரி அல்லது தவறு என்ற புலனுணர்வுடன் உண்மையில் இணங்கவேண்டும் என்பதில்லை.

7.3.3. விதி அடிப்படையிலான திருத்தம் (Rule based checking)

இது டேனியன் ராபர் மற்றும் பிறரால் (Denial Naber et al) பயன்படுத்தப்பட்ட அணுகுமுறையில் குறைந்தது சொல்வகைப்பாட்டுக்காக அடையாளப் படுத்தப்பட்ட உரை ஒரு குழும விதிகளாகப் பொருத்தம் பார்க்கப்படும். இந்த அணுகுமுறை புள்ளியியல் அடிப்படையிலான அணுகுமுறைக்கு ஒற்றுமையுள்ளது; ஆனால் உரைகள் மனிதனால் உருவாக்கப்பட்டதாகும்.

பிற அணுகுமுறைகளில் கூறப்பட்ட சிக்கல்கள் காரணமாகக் கண்டிப்பான விதி அடிப்படையிலான ஒழுங்குமுறை இங்கு உருவாக்கப்படும். தொடரியல் அடிப்படையிலான பிழை திருத்திகளைப் போல அல்லாமல் விதி அடிப்படையிலான பிழை திருத்தி ஒரு போதும் முழுமையாக இருப்பதில்லை; அதாவது அதில் கண்டுபிடிக்க இயலாத பிழைகள் இருந்துகொண்டே இருக்கும். இருப்பினும் அதற்கு பல அனுகூலங்கள் உண்டு:

- ஒரு வாக்கியம் சரிபார்க்க முழுமையாக இருக்கத் தேவையில்லை; மாறாக மென்பொருள் பனுவலை அது தட்டச்சு செய்யப்படும் போதே சரிபார்த்து உடனடியான பதில் கருத்தைத் தரவியலும்.
- ஒவ்வொரு விதியும் வெளிப்பாட்டு விளக்கம் கொண்டிருப்பதால் அதற்கு உருவங்கொடுப்பது எளிது மற்றும் தனித்தனியாக முடுக்கவோ அணைக்கவோ இயலும்.
- அது பயனுள்ள கருத்துரைகளுடன் இலக்கண விதிகளை விளக்கியே விளக்கமான பிழைச் செய்திகளைத் தர இயலும்.
- விதி ஒழுங்கமைப்பு குறைந்தது பல எளிய ஆனால் பொதுவான பிழை நேர்வுகளை எளிதில் புரிந்துகொள்ள இயலுமாகையால் அதை அதன் பயன்பாட்டாளர்களுக்கு எளிதாக நீட்சி செய்யவியலும்
- அதைப் படிப்படியாக ஒரு விதியில் தொடங்கி பின்னர் விதிக்கு விதி நீட்சி செய்து உருவாக்கவியலும்.

7.4. நடை மற்றும் இலக்கணத்திருத்தி

நடை மற்றும் இலக்கண திருத்தி (style and grammar checker) ஒரு பனுவலை/உரையை ஏற்கும் பின்னர் சாத்தியமான தவறுகளைத் திருப்பி அனுப்பும். தவறுகளை/பிழைகளைக் கண்டுபிடிப்பதற்காக பனுவலின் ஒவ்வொரு சொல்லுக்கும் அதன் சொல்வகைப்பாடு மதிப்பிடப்படும்/தரப்படும்; ஒவ்வொரு வாக்கியமும் தொடர் உறுப்புகளாக/ தொடர்பன்களாகப் (chunkers) பிரிக்கப்படும். எடுத்துகாட்டாகப் பெயர்த்தொடர்கள் பின்னர் பனுவல் பிழை திருத்தியின் முன்வரையறை செய்யப்பட்ட பிழை விதிகளுடன் ஒப்பிட்டுப்பார்க்கப்படும். ஒரு

விதி பொருந்தினால் பொருந்தும் இடத்தில் உரையில் பிழை இருப்பதாகக் ஊகிக்கப்படும்/கருதப்படும். விதிகள் பிழைகளைச் சொற்கள், சொல்வகைப்பாட்டு அடையாளங்கள், தொடர்பன்கள்/தொடர் உறுப்புகள் இவற்றின் ஒருங்கமைப்பாக (pattern) விளக்கும். ஒவ்வொரு விதியும் பிழைக்கு விளக்கத்தை உட்படுத்தும்; அது பயன்படுத்துபவர்க்குக் காட்டப்படும்.

அமைப்பொழுங்கின் விதி அடிப்படையிலான அணுகுமுறை பயன்படுத்துபவர்கள் தங்களது விதிகளை உருவாக்கிக்கொள்ள இயலச்செய்கின்ற அளவுக்கு எளிமையானதாகும்; பல எடுத்துக்காட்டான பிழைகளைக் கண்டுகொள்வதற்குப் போதுமான அளவு திறன் வாய்ந்ததாகும். விதிகள் எளிய XML வடிவமைப்பில் வெளிப்படுத்தப்படுகின்றது; இது பிழைகளை விளக்குவதுடன் பயன்படக்கூடிய பிழைச் செய்திகளையும் எடுத்துக்காட்டு வாக்கியங்களையும் கொண்டிருக்கின்றது.

7.4.1. கோட்பாட்டுப்பின்னணி

எழுத்துப்பிழை திருத்தம் பயனுள்ளதாகக் கருதப்பட்ட அதே காரணத்திற்காக நடை மற்றும் இலக்கணப்பிழை திருத்தமும் பயனுள்ளதாகும். இது மக்கள் பிழையில்லாமல் ஆவணங்களை எழுதுவதற்குப் பயன்படுகின்றது. நடை மற்றும் இலக்கணப்பிழை திருத்தி பயனுள்ளதாக அமையச் சில தேவைகளைப் பூர்த்திசெய்யவேண்டும்.

- அது விரைவாக இருக்கவேண்டும்; அதாவது இடையீடுப் பயன்பாட்டிற்குப் போதுமான அளவு விரைவாக இருக்கவேண்டும்.
- அது இருக்கின்ற சொல்லாய்வியுடன் நன்றாக இணைக்கப் பட்டிருக்கவேண்டும்.
- சரியான வாக்கியங்களைப் பெரும்பான்மையும் தவறாகக் காட்டக்கூடாது.
- சொந்தத் தேவைகளுக்காக அதை மாற்றிக்கொள்வது சாத்தியப்படவேண்டும்.
- இறுதியாக அது முழுமையாக இருக்கவேண்டும்; அதாவது அது பனுவலில் உள்ள பெரும்பான்மையான பிழைகளைக் கண்டுபிடிக்கவேணடும்.

எழுத்துப்பனுவலில் தோன்றும் பிழைகளைப் பல்வேறு பட்ட வழிகளில் வகைப்படுத்த இயலும்.

எழுத்துப் பிழைகள்: இது எழுத்துப்பிழை திருத்தியால் கண்டுபிடிக்க இயலும் பிழையாகும் என வரையறை விளக்கம் செய்யப்படுகின்றது. இலக்கணப் பிழைதிருத்திகள் அறியப்பட்ட சொற்களின் பெரிய பட்டியலுடன் சொற்களை ஒப்பிட்டுப்பார்க்கும். சொல் அந்தப் பட்டியலில் இல்லாவிட்டால் அது பிழையானது என்று கருதப்படும். ஒற்றுமையுள்ள (சரியான) சொற்கள் பதில் சொற்களாகத் தரப்படும். எடுத்துக்காட்டாக *Gemran என்பது German எனத் திருத்தப்படும்.

இலக்கணப் பிழைகள்: ஒரு மொழியின் இலக்கண விதிகளுடன் ஒத்துப்போகாத ஒரு வாக்கியத்தை விளைவிக்கும் பிழை இலக்கணப் பிழையாகும். எழுத்துப்பிழை திருத்தம் போல அல்லாமல் இலக்கணப்பிழை திருத்தம் சூழல் ததகவலைப் பயன்படுத்த வேண்டியது அவசியமாகும்.

*Harry Potter bigger then than Titanic?

இப்பிழை தட்டச்சால் ஏற்பட்டதா, எழுதுபவரின் மனதில் *then* என்பதற்கும் than என்பதற்கும் இடையில் ஏற்பட்ட தவறான புரிந்துகொள்கையால் ஏற்பட்டதா என்பதைத் தீர்மானிக்க இயலாது. *then* என்பதைப் பயன்படுத்துவது தவறு என்று தெளிவாகத் தெரிவதால் இது இலக்கணப் பிழையாகும்.

இலக்கணப் பிழைகளை அமைப்புசார் பிழை மற்றும் அமைப்புசாரா பிழை என்று பகுக்கலாம். அமைப்பு சார்ந்த பிழைகளைச் சொற்களைச் செருகியோ நீக்கியோ ஒன்றோ இரண்டோ சொற்களை நகர்த்தியோ மட்டுமே திருத்த இயலும். அமைப்புசாராத பிழைகளை இருக்கிற சொல்லை வேற்று சொல்லால் இடம்பெயர்த்து திருத்தம் செய்ய இயலும்.

நடைப்பிழைகள்: இயல்பற்ற/அசாதாராண சொற்களும் சிக்கலான வாக்கிய அமைப்புகளும் உரையை/பனுவலைப் புரிந்துகொள்வதைக் கடினமாக்குகின்றது; இது யாராலும் விரும்பப்படுவதில்லை. இந்நேர்வுகள் நடைப் பிழைகள் என்று கருதப்படும். இலக்கணப்பிழைகளைப் போல் அல்லாமல் நடைப் பிழைகள் அதிகமாகச் சூழலைச் சார்ந்தும் பனுவல்/உரை வகையைச் சார்ந்தும் அமையும். எடுத்துக்காட்டாக, சொந்த நண்பர்களுக்கிடையிலான மின்னஞ்சல் கருதுப்பரிமாற்றம் மொழியின் படைப்பாக்கப் பயன்பாட்டை அனுமதிக்கின்றது; அனால் தொழில் நுட்ப ஆணவப்படுத்தல் பொருண்மை மயக்கத்தால் பாதிக்கலாகாது. அமைப்பு உறுதிப்பாடு (configurability) இலக்கணப் பிழை திருத்தங்களை விட நடை பிழை திருத்தத்திற்குக் கூடுதல் முக்கியத்துவம் வாய்ந்ததாகும்.

எடுத்துக்காட்டாக, பினவரும் வாக்கியத்தை எடுத்துக்கொள்ளவும்.

> But it [a human reason] quickly discover that, in this way, its labours must remain ever incomplete, because new questions never cease to present themselves, and thus it finds itself compelled to have recourse to principles which transcend the region of experiences, while they are regarded by common sense without distrust.

இவ்வாக்கியம் உண்மையான காரணத்தால் காந்தின் விமர்சனத் திறனாய்விலிருந்து (Kant's Critique) விளைந்ததாகும். இது 48 சொல் நீளமுள்ளதாகும்; பலரும் இது புரிந்துகொள்வதற்குக் கடினமானது என்று ஒத்துக்கொள்வர். இதற்குக் காரணம் அதன் நீளம், (transcend போன்ற) கடினமான சொற்றொகை, இரட்டை எதிர்மறையின் (double negation) பயன்பாடு.

இக்காலகட்டத்தில் புரிந்துகொள்வதற்கு எளிமையான ஆணவத்தின் தேவையுள்ள சூழலில், இவ்வாக்கியத்தில் நடைச் சிக்கல் உள்ளதாகக் கருதலாம்.

பொருண்மைப் பிழைகள்: நடைப் பிழையோ இலக்கணப்பிழையோ எழுத்துப் பிழையோ இல்லாத தவறான தகவல் உள்ள ஒரு வாக்கியத்தில் பொருண்மைப் பிழை உள்ளதாகக் கருதலாம். பொருண்மைப் பிழைகளைக் கண்டுபிடிக்க மிகப் பரந்த உலகறிவு (world knowledge) தேவையாதலால், இப்பிழைகளை வழக்கமாகக் கண்டுபிடிக்க இயலாது. எடுத்துக்காட்டாக, "MySQL is a great editor for programming!" என்ற வாக்கியம் உண்மையோ பொய்யோ அல்ல; MySQL என்பது ஒரு எடிட்டர் அல்ல ஒரு டேட்டா பேஸ் ஆகும் என்பதால் இது பொருளைத் தரவில்லை. பரந்த உலகறிவு இல்லாமல் இதைத் தெரிந்துகொள்ள இயலாது. உலகறிவு சூழலின் ஒரு வடிவாகும்; ஆனால் இன்றைய காலகட்டத்தில் இது மென்பொருளின் புரிந்துகொள்ளலுக்கு அப்பாற்பட்டதாகும்.

பிழைகளுக்கும் (errors) தவறுகளுக்கும் (mistakes) வேறுபாடு பாராட்டவேண்டியதில்லை. தவறாகப் கருதப்படுபவைகளும் மோசமாக எழுதப்படுபவைகளும் பிழைகளாகும். இலக்கணம் (அல்லது தொடரியல்) சரியான வாக்கியங்கள் எவ்வாறு இருக்கவேண்டும் என்பதை விளக்கும் விதிகளின் ஒழுங்கமைப்பாகும் (system of rules). எவ்வாறோ விதிகள் மக்கள் மனதில் இருக்கின்றது; இதன் காரணமாகப் பெரும்பாலான மக்கள் ஒரு வாக்கியம் தவறானதா இல்லையா என்பதை எளிதில் தீர்மானிக்க இயலும். உள்ளுணர்வு / மன உணர்வு (intuition) அடிப்படையில் சரியானது/தவறானது என்ற தீர்மானத்தைக் கடினமாக்கும் எல்லையை ஒட்டிவரும் நேர்வுகளைக் காண்பது சாத்தியமாகும். பிழையான செய்திக்குப் பதிலாக விளக்க எச்சரிக்கைச் செய்தியைக் (descriptive warning message) காட்டி ஒரு மென்பொருள் இந்நேர்வுகளைக் கையாளும். இலக்கண விதி சில சரியான வாக்கியங்களுக்கும் பொருந்துவது தெரிந்தால், எச்சரிக்கைச் செய்திகளும் நடைப் பிழைகளுக்கும் இலக்கணப் பிழைகளுக்கும் பயனுள்ளவையாக அமையும்.

7.5. தமிழில் எழுத்துப்பிழை மற்றும் இலக்கணப்பிழைத் திருத்தி உருவாக்கம்

இங்கு தமிழுக்கென்று உருவாக்கப்பட்ட எழத்துப் பிழை மற்றும் இலக்கணப் பிழை திருத்திகளின் சுருக்கமான தகவல்கள் தரப்படும். மேலும் இவ்வாய்வின் விளைவாக உருவாக்கப்பட்ட எழுத்துப் பிழை மற்றும் இலக்கணப் பிழை திருத்தியின் மாதிரி விளக்கப்படும். தமிழுக்கென்று எழுத்துப்பிழை திருத்தி உவாக்கும் முயற்சிகள் பல மேற்கொள்ளப்பட்டுள்ளன. சில சொல்லாய்விகளில் இத்தகைய எழுத்துப்பிழை திருத்திகள் பயன்படுத்தப்படுகின்றன. இருப்பினும் முழுமையான எழுத்துப் பிழை திருத்தி இதுவரை உருவாக்கப் படவில்லை என்றே

கூறலாம். இதுவரை உருவாக்கப்பட்ட எழுத்துப் பிழை திருத்திகள் பற்றி இங்குச் சுருக்கமாக விளக்கப்பட்டுள்ளது.

7.5.1 தமிழ்ப் பல்கலைக்கழகப் பிழைத் திருத்தி

தமிழ்ப் பல்கலைக் கழக மொழியியல் துறையில் இராசேந்திரன் அவர்களால் 1999-2001 காலகட்டத்தில் தமிழுக்கு எழுத்துப் பிழை திருத்தியும் இலக்கணப்பிழை திருத்தியும் உருவாக்கும் முயற்சி பல்கலைக்கழக மான்யக்குழுவின் நிதிநல்கையால் மேற்கொள்ளப்பட்டது (Rajendran 1999). இது உருபனியல் பகுப்பாய்வி அடிப்படையில் அமைந்திருந்தது. உருபனியல் பகுப்பாய்வி அனுசாரகா திட்டம் அடிப்படையில் அமைந்திருந்தது. இது ஒரு அடுக்குமுறை அணுகுமுறையாகும் (paradigmatic approach). இதில் வினைகளும் பெயர்களும் அவைகளின் ஒலியனியல்-உருபனியல் நடத்தை அடிப்படையில் குழுக்களாகப் பிரிக்கப்பட்டு அதன் அடிப்படையில் வினைவடிவங்களும் பெயர்வடிவங்களும் ஆயப்பட்டன. தவறான உருபனியல் பகுப்பாய்வு விளைந்தால் அச்சொல் தவறு என்று சுட்டிக்காட்டப்படும். சரியன உருபனியல் பகுப்பாய்வு விளைந்தால் சொற்கள் சரியானவை என்று சுட்டிக்காட்டப்படும். இந்த அளவில் தான் இவ்வாய்வு அமைந்தது. இதற்கு மேல் இவ்வாய்வு மேற்கொள்ளப்படவில்லை. இருப்பினும் தமிழின் ஒலியனியல் அமைப்பும் உருபனியல் அமைப்பும் ஆயப்பட்டு தகவல்கள் திரட்டப்பட்டன. ஒலியன் வருகைமுறை விதிகள், உருபன் வருகை முறை விதிகள், சொற்கள் மற்றும் தொடர்கள் வருகை முறை விதிகள் யாவும் திரட்டப்பட்டு கணினி மொழியியல் ஆய்வுக்காக பட்டியலிடப்பட்டன. சொற்களின் வகைப்பாடு அடையாளப்படுத்தப்பட்ட ஒரு மின் அகராதியும் உருவாக்கப்பட்டது.

7.5. 2. தனபாலன் மற்றும் பிறரின் தமிழ் எழுத்துப்பிழைத் திருத்தி

தனபாலன் மற்றும் பிறர் தமிழுக்கு என்று உருவாக்கிய எழுத்துப் பிழை திருத்தி (Dhanabalan, Ranjani Parthasarathi and Geetha) அகராதிகளைப் பயன்படுத்துகின்றது. உருபனியல் மற்றும் தொடரியல் தகவல்களைக் கொண்ட அகராதிகள் சொல் ஆய்விகளில் ஒருங்கிணைக்க இயலும் தமிழ் எழுத்துப்பிழை திருத்தி உருவாக்கவும் அதிக கலவைத் தன்மையான இயற்கைமொழி ஆய்வுப் பயன்பாடுகளில் பயன்படுத்த இயலும் உருபனியல் மற்றும் தொடரியல் ஆய்விகளின் உருவாக்கத்திற்கும் தேவைப்படுகின்றது.

7.5.2.1. தமிழ்ப் பிழைத்திருத்தத்தின் சிக்கல்கள்

இவர்கள் பிழை திருத்தத்தின் முக்கியச் சிக்கல்களாகப் பின்வருவனவற்றைப் பட்டியலிடுகின்றனர்.

1. வேற்றுமை உருபுகளைக் கையாளுதல்: ஆங்கிலத்தில் தனிச் சொல்லாக வரும் முன்னுருபுகள் தமிழில் வேற்றுமை உருபுகளாகவும் (பின்னொட்டுகளாகவும்) முன்னுருபுகளாகவும் வரும்.

   எ.கா. ராமன் கிருஷ்ணனைப் பற்றிச் சொல்கின்றான்.

2. வேற்றுமை உருபுகளால் வேர்ச்சொல்லைப் பெயராகவோ வினையாகவோ வகைப்படுத்துதல்.

3. குற்றுயிர்-நெட்டுயிர் பிழைகளைக் கையாளுதல்.

கிணறு - கேணறு

4. இரு சொற்களுக்கு இடையில் வரும் புணர்ச்சி விதிகளைக் கையாளுதல்.

அவனுக்குப் பிறகு

7.5.2.2 தமிழ்ப் பிழைத்திருத்தியின் தொகுதிகள்

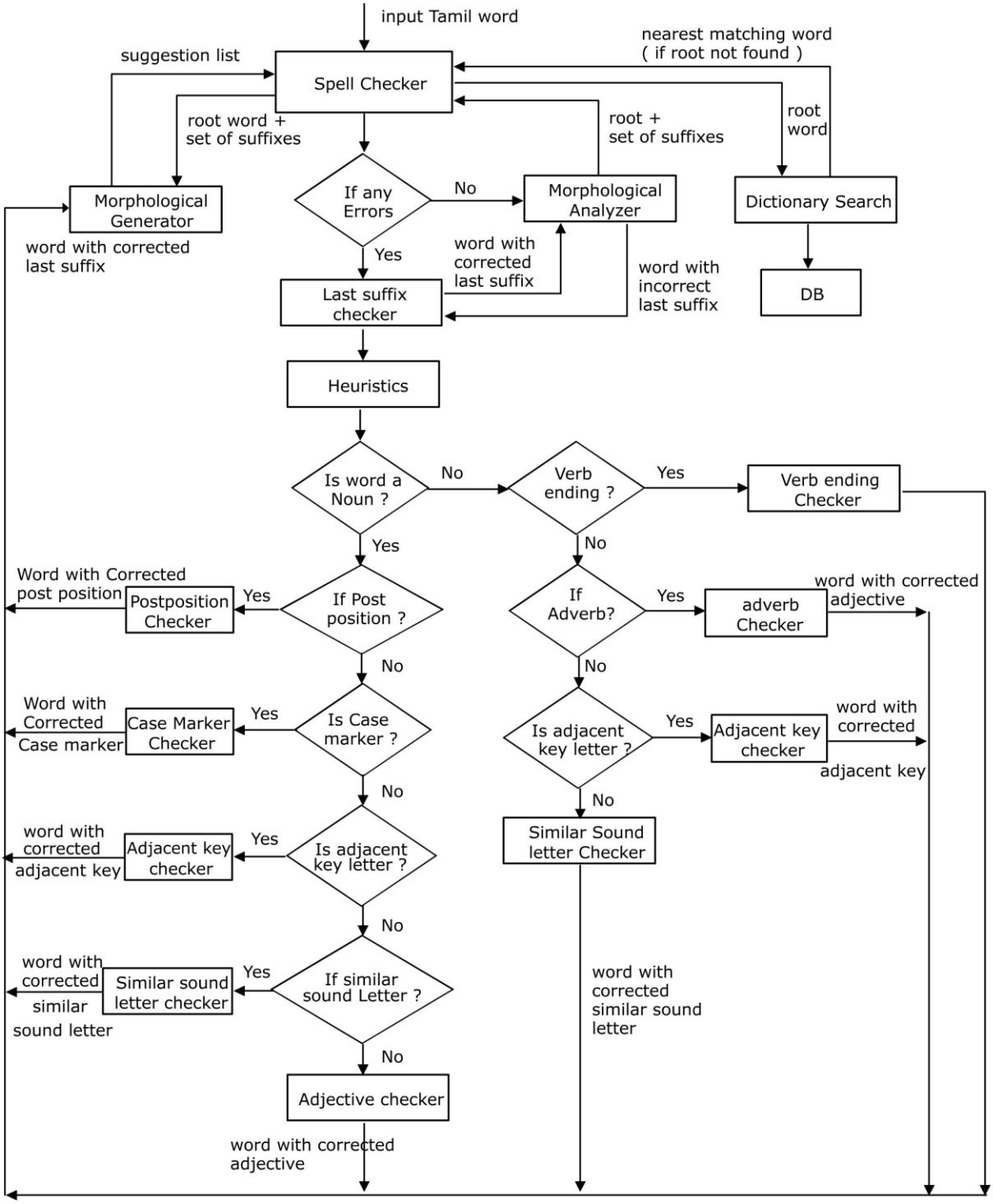
தமிழ் பிழை திருத்தியின் பொதுவான அமைப்பு ஒழுக்குப் படத்தில் தரப்பட்டுள்ளது. தொடக்கத்தில் பிழைதிருத்தி ஆவணத்திலிருந்து எடுக்கப்பட்ட சொற்களை ஒரு வேளைக்கு ஒன்று என்று படிக்கின்றது. அகராதி பின்னர் கிடைத்த சொல்லை பரிசோதிக்கின்றது. சொல் அகராதியில் காணப்பட்டால் அது ஒரு சரியான சொல்லாகக் கருதப்படும். சொல் அகரதியில் காணப்படாவிட்டால் அது பிழைதிருத்தத்திற்கு அனுப்பப்படும். பிழை திருத்தி பின்வரும் பகுதிகளைக் கொண்டது:

- பனுவல் பகுத்துக் குறித்தல்
- எழுத்துக் கூட்டல் பரிசோதனை மற்றும் திருத்தல்
- பரிந்துரைப் பட்டியல் உருவாக்கம்

இப்பகுதிகளுக்கு உதவப் பிழை திருத்தி பின்வருவனவற்றைப பயன்படுத்தும்.

- தரப்பட்ட சொல்லை ஆயும் உருபனியல் ஆய்வி
- பரிந்துரைகளை உருவாக்கும் உருபனியல் உருவாக்கி

பனுவல் பகுத்துக் குறித்தல் உள்ளீடு செய்யப்பட்ட ஆவணத்தில் பிழை திருத்துவதற்குப் பொருத்தமான ஒரு சொல்லை அதைச் சுற்றியுள்ள பனுவலிலிருந்து தனியாகப் பிரிக்கின்றது. கண்டுபிடிக்கப் பட்ட சொல் பிழையானது என்றால் அது பின்னர் எழுத்துப்பிழை திருத்தப் பகுதிக்கும் பரிந்துரைகள் உருவாக்கப் பகுதிக்கு அனுப்பபடும். ஆவணங்களச் சொற்களின் பட்டியலாகப் பகுத்துக்குறித்த பின்னர் ஒவ்வொரு சொல்லும் உருபனியல் பகுப்பாய்விக்கு (morphological analyzer) அனுப்பப்படும். உருபனியல் பகுப்பாய்வி பின்னொட்டைப் பிரிக்க முயலும். அது சரியான சொற்களை மட்டும் பகுத்தாயும் படிக்கு திட்டமிடப்பட்டுள்ளது. எழுத்துப்பிழை காரணமாக அது பின்னொட்டைப் பிரிக்க இயலவில்லை என்றால் அது அச்சொல்லை எழுத்துக் கூட்டல் பரிசோதனை மற்றும் திருத்தப் பகுதிக்கு அனுப்பும். சொல்லை வேர்களாகவும் பின்னொட்டுக்களாகவும் பிரித்த பின்னர் அகராதியுடன் ஒப்பிடப்பட்டு அதன் செல்லுபடி சரிபார்க்கப்படும். அச்சொல் அகராதியில் இல்லை என்றால் அண்மை அடிப்படையில் பொருந்துகிற பரிந்துரைகள் தெரிந்தெடுக்கப்படும்.



7.5.2..2.1 பெயர்களைச் சரிபார்த்தல்

தமிழில் பெயர் ஒரு சொல்லாகும். இது வேற்றுமை உருபுகளையும் பன்மை உருபுகளையும் பின்னுருபுகளையும் குறைசொல் (clitics) பின்னொட்டுக்களையும் எடுக்க வல்லன. பெயர் சொல்பகுதிகளை ஒருபன் சொற்கள் (எ.கா. *கல், மலர்*) என்றும் பல்லுருபன் சொற்கள் (எ.கா. *கல்வி, காட்சி*) என்றும் பகுக்க இயலும். ஒருபன் சொற்களை வேர்களாகவும் ஒட்டுக்களாகவும் பிரிக்க இயலாது; ஆனால் பல்லுருபன் சொற்களை வேர்களாகவும் ஒட்டுக்களாகவும் பிரிக்க இயலும். வேற்றுமை உருபுகள் பெயர்த்தொடருக்கும் வினைப் பயனிலைக்கும் இடையே உள்ள தொடரியல் மற்றும் பொருண்மை உறவுகளை வெளிப்படுத்தும். தமிழில் பெரும்பாலான வேற்றுமைகள் பின்னொட்டுக்களாலும் சில வேற்றுமை பின்னொட்டுக்களுடன் இணைந்த பின்னுருபுகளாலும் மெய்ப்படுத்தம் செய்யப்படும். இங்கு பத்து வேற்றுமைகள் கருத்தில் கொள்ளப்பட்டுள்ளன; அவற்றின் பட்டியல் பின்னொடுக்களுடன் அட்டவணை 1-இல் தரப்பட்டுள்ளன.

வ.எண்	வேற்றுமை	பின்னொட்டு
1	எழுவாய் வேற்றுமை	0
2	செயப்படுபொருள் வேற்றுமை	ஐ
3	கொடை வேற்றுமை	கு
4	பயன் வேற்றுமை	உக்காக
5	கருவி வேற்றுமை	ஆல்
6	உடனிகழ்ச்சி வேற்றுமை	ஆல்
7	இட வேற்றுமை	இல்
8	நீங்கல் வேற்றுமை	இல்லிருந்து, இடமிருந்து
9	கிழமை வேற்றுமை	அது

பிழைதிருத்தச் செயல்பாடு வேற்றுமை உருபு சரிபார்த்தல், பின்னுரு சரிபார்த்தல், பெயரடை சரிபார்த்தல் என்று மேலும் பகுக்கப்படும். *இல், ஆல், இடம், உடன், ஆக, கல், உக்கு, ஓடு, இருந்து, உடைய, அது* போன்ற வேறுபட்ட வேற்றுமை உருபுகள் உருபனியல் பகுப்பாய்வியால் பிரித்தெடுக்கப்படுகின்றன. இக்கட்டத்திற்குப் பின்னர் இறுதி ஒட்டின் செல்லுபடியைச் தீர்மானிக்க ஒரு சரிபார்ப்பு நடத்தப்பெறும். பின்வருவன பெயர்களில் வேற்றுமை உருபுகளில் விளையும் சில பிழைகளாகும்.

எ.கா.

மரத்தில் என்பதற்குப் பதிலாக *மரத்தல்* என வருதல்

அவனிடம் என்பதற்குப் பதிலாக *அவனேடம்* என வருதல்
அவனுக்காக என்பதற்குப் பதிலாக *அவனுக்க* என வருதல்
மரங்கள் என்பதற்குப் பதிலாக *மரங்காள்* என வருதல்
மரத்திலிருந்து என்பதற்குப் பதிலாக *மரத்திலருந்து* என வருதல்
மரத்துக்கு என்பதற்குப் பதிலாக *மரத்தக்கு* என வருதல்

மரத்திலருந்து என்ற சொல்லை எழுத்துப்பிழை சரிபார்க்க எடுத்துக்கொள்ளவும். இச்சொல் பகுப்பாய்வுக்காக உருபனியல் பகுப்பாய்விக்கு அனுப்பப்படும். சொல்லின் பின்னொட்டு *இலருந்து* என்று இருப்பதன் காரணமாக உருபனியல் பகுப்பாய்வியால் இச்சொல்லைப் பிரிக்க இயலவில்லை. எழுத்துக்கூட்டல் சரிபார்ப்பும் பிழைதிருத்தக் கட்டமும் இறுதி பின்னொட்டை (*இலருந்து* என்பதை) *இலிருந்து* எனத் திருத்தி சொல்லைப் பிரிப்பதற்காக உருபனியல் பகுப்பாய்விக்கு அனுப்பும். இறுதியில் அது வேர்ச்சொல்லையும் பினொட்டையும் பரிந்துரை உருவாக்கக் கட்டத்திற்கு அனுப்பும்.

7.5.2.2.2 வினைச் சொற்களைச் சரிபார்த்தல்

கால ஒட்டுக்களை ஏற்கவல்ல சொல் வினையாகும். வினைச் சொல் வினையடையால் அல்லது வினையெச்சத் தொடரால் அடைசெய்யப்படலாம்; வினை இடம்-எண்-பால் ஒட்டுக்களால் திரிபுறலாம் மற்றும் பெயரெச்சமாகவோ வினை எச்சமாகவோ மாற்றப்படலாம். வினைகள் இரு குழுமங்களாகப் பகுக்கப்படும்: ஒன்று முற்று வடிவம்; மற்றொன்று எச்சவடிவம்.

முற்றுவடிவம்

வினையின் முற்றுவடிவம் தனிநிலை வடிவாகவோ கூட்டுநிலை வடிவாகவோ இருக்கலாம். தனிநிலை முற்று வடிவங்கள் அமைப்பு வகுப்புகளின் படி மூன்று குழுமங்களாகப் பகுக்கப்படும்.

1. பின்னொட்டுகள் எடுக்காத வினைகள். எ.கா. *சரி, தனி.*
2. இடம்-எண்-பால் ஒட்டு ஏற்காத காலமில்லா வினைகள். எ.கா. *இல்லை, உண்டு*
3. இடம்-எண்-பால் ஒட்டுக்களால் திரிபுறும் மற்றும பிற ஒட்டுக்களையும் ஏற்கும் வினைகள். எ.கா. *சாப்பிடு*

வினையடிகள் அவைகள் ஏற்கும் கால ஒட்டுகள் அடிப்படையயில் வகைப்படுத்தப் படவேண்டும்.

தனிநிலை முற்றுவினைகள்

கால ஒட்டுகளையும் இடம்-எண்-பால் ஒட்டுக்களையும் ஏற்கும் வினைகள் தனிநிலை வினைகள் என அழைக்கப்பெறும். இதில் வேர்களைத் தொடர்ந்து கால ஒட்டுகளும் அதனைத் தொடர்ந்து இடம்-எண்-பால் ஒட்டுக்களும் சேர்க்கப்படும். மனித எழுவாய்களுக்கும் மனிதரல்லாத எழுவாய்க்களுக்கும் கால ஒட்டுகள் எப்பொழுதும் ஒன்றாக இருப்தில்லை. மேலும் மனிதரல்லாத

ஒருமை எழுவாய்க்கும் மனிதரல்லாத பன்மை எழுவாய்க்கும் எப்பொழுதும் கால ஒட்டுகள் ஒன்றாக இருப்பதில்லை. இடம்-எண்-பால் ஒட்டுக்கள் எழுவாயுடன் உடன்பாடு கொண்டிருக்கும். கால ஒட்டுக்குப் பதிலாக எதிர்மறை ஒட்டான ஆ என்பதையும் வினைவேருடன் நேரடியாகச் சேர்த்து எதிர்மறை வினைவடிவைப் பெறவியலும்.

எ.கா.

செய்-கிற்-ஆன் (நிகழ்காலம்)

செய்-த்-ஆன் (இறந்தகாலம்)

செய்-வ்.ஆன் (எதிர்காலம்)

செய்ய்-ஆ-து (எதிர்மறை வடிவம்)

கூட்டுநிலை முற்று வினைகள்

கூடுநிலை வினை வினையடி, வினையாறுவகை, வினைப்பாடு, வினைநோக்கு, காலம், இடம்-எண்-பால் மற்றும் குறைச்சொல் இவற்றைக் கொண்டிருக்கும். இவற்றில் வினையடியைத் தொடரும் ஒட்டுகள் விருப்பாகும். கூட்டுநிலை வினைகள் ஒன்றோ அதற்கு மேலோ துணைவினைகளைக் கொண்டிருக்கும். பொதுவாக ஒருவகைத் துணைவினனை அதேபோன்ற துணைவினையைத் தொடராது.

வினையின் எச்ச வடிவங்கள்

வினை எச்ச வடிவங்களைப் பெயரெச்ச வடிவம், வினையெச்ச வடிவம் என்று வகைப்படுத்தலாம். பெயரெச்ச வடிவைக் கால ஒட்டுடன் *அ* என்ற ஒட்டைச் சேர்த்துப் பெறலாம்.

எ.கா.

படி-த்த்-அ > படித்த

பதி-க்கின்ற்-அ > படிக்கின்ற

படி-உம் > படிக்கும்

வினையுடன் *அ* அல்லது *து* அல்லது *ஆல்* என்ற ஒட்டை நேரடியாகச் சேர்த்து வினையெச்ச வடிவைப் பெறலாம்.

எ.கா.

போ-க > போக (செய்ய வினையெச்சம்)

படி-க்க > படிக்க (செய்ய வினைஎச்சம்)

பார்-த்து > பார்த்து (செய்து வினையெச்சம்)

செய்-த்-ஆல் > செய்தால் (நிபந்தனை வினையெச்சம்)

படி-த்-ஆல் > படித்தால் (நிபந்தனை வினையெச்சம்)

வினை சரிபார்ப்பு இடம்-எண்-பால் சரிபார்ப்பு, கால ஒட்டு சரிபார்ப்பு என மேலும் வகைபடுத்தப்படும். வினையின் வேறுபட்ட ஒட்டுகள் உருபனியல் ஆய்வியால் பிரித்தெடுக்கப்படும். கால ஒட்டுகள் *கிற், கின்று, ஆனின்று, ப், வ், த், ட், ற்* என்பனவற்றை உட்படுத்தும். இவை வினைகளை நிகழ்காலம் அல்லது இறந்தகாலம் அல்லது எதிர்காலம் என்று அடையாளப்படுத்தும். இடம்-எண்-பால் ஒட்டுகள் *அன், ஆன், அள், ஆள், அர், ஆர், து, அ, வை, கள்* என்பனவற்றை உட்படுத்தும். வினையின் ஒட்டுகளில் சொல் பிழையாக எழுதப்பட்டால் பிழை திருத்தப்பட்டு சாத்தியமான பரிந்துரைகள் உருவாக்கப்படும்.

எ.கா.

செய்கிறார்கள் என்பதற்குப் பதிலாகச் *செய்கிறர்கள்* என வருதல்
செய்தது என்பதற்குப் பதிலாகச் *செய்தத* என வருதல்
செய்வான் என்பதற்குப் பதிலாகச் *செய்வன்* என வருதல்

தோன்றுகின்றன் என்ற சொல்லைக் கருத்தில் கொள்ளவும். இச்சொல் பகுத்தாயப்படுவதற்காக உருபனியல் பகுப்பாய்வியிடம் அனுப்பப்படுகின்றது. எழுத்துக்கூட்டல் சரிபார்ப்பு, திருத்தம் என்ற கட்டங்கள் *தோன்றுகின்றன்* என்ற சொல்லின் இறுதி ஒட்டான *அன்* என்பதை *ஆன்* என திருத்துகின்றது. பகுப்பாய்வி வேர்ச் சொல்லையும் பின்னொட்டுக்களையும் பரிந்துரை உருவாகக்கட்டத்திற்கு அனுப்புகின்றது. உருபனியல் பகுப்பாய்வியால் அனுப்பப்பட்ட வேர்ச் சொல்லையும் பின்னொட்டுகளையும் பரிந்துரை உருவக்கக் கட்டம் பரிந்துரை உருவாக்கத்திற்காக உருபனியல் உருவாக்கியிடம் அனுப்புகின்றது.

7.5.2.2.3 பின்னுருபுகளைச் சரிபார்த்தல்

வேற்றுமை உருபின் செயல்பாட்டைச் செய்யும் சொல் பின்னுருபு என அழைக்கப்படும். பொதுவாக பின்னுருபு வேற்றுமை பின்னொட்டிற்குப் பின்னர் வரும். சில இடங்களில் வேற்றுமை பின்னொட்டு விருப்பாக வரும். பின்னுருபுகள் அவற்றின் முன்வரும் வேற்றுமை உருபுகளின் அடிப்படையில் ஐந்து வகுப்புகளாக வகைபடுத்தப்படுகின்றன.

செயப்படுபொருள் வேற்றுமையின் பின்னர் வரும் பின்னொட்டுகள்:

விட, போல, கொண்டு, பற்றி, குறித்து, சுற்றி, விட்டு, தவிர, முன்னிட்டு, வேண்டி, ஒட்டி, பொறுத்து, பொறுத்தவரை, …

உடைமை வேற்றுமையின் பின்னர் வரும் பின்னொட்டுகள்:

மீது, மேல், வழியாக, மூலமாக, வாயிலாக, கீழ், …

இட வேற்றுமையின் பின்னர் வரும் பின்னொட்டுகள்:

இருந்து, உள்ள

கொடை வேற்றுமையின் பின்னர் வரும் பின்னொட்டுகள்:

ஆக, என்று, முன், பின், உரிய, உடைய, மத்தியில், எதிரில், அருகில், பதில், தகுந்த, மாறாக, நேராக, …

உள்ளீட்டுச் சொல் ஒரு பின்னொட்டாக இருந்தால் அது சரியா தவறா என்று சரிபார்க்கும். சொல் தவறு என்றால் நேராக்கப்படும் மற்றும் பரிந்துரைகள்/ஆலோசனைகள் உருவாக்கப்படும்.

7.5.2.2.4. ஒன்றுபோல் உச்சரிக்கப்படும் எழுத்துக்கள்

ஒன்றுபோல் உச்சரிக்கப்படும் எழுத்து சொல்லின் எழுத்துமுறையில் தவறை விளைவிக்கக்கூடும். எடுத்துக்காட்டாகத் *தாளம்* என்ற சொலை எடுத்துக்கொள்வும். *ள* என்ற எழுத்து *ல* எனத் தவறாக (தாலம்) எழுதப்படலாம். தமிழில் மூன்று எழுத்துக்களுக்கு அவற்றைப் போலவே உச்சரிக்கப்படும் எழுத்துக்கள் உள்ளன. தவறான சொல்லுக்குச் சாத்தியமான ஒன்றுபோல் உச்சரிக்கப்படும் எழத்துக்களைப் பரிசோதித்து அகராதியில் உள்ள பரிந்துரைகள் பயன்படுத்துவோருக்குக் காட்டப்படும். தமிழில் ஒன்றுபோல் உச்சரிக்கப்படுகின்ற சாத்தியமான எழுத்துக்கள் பின்வருவனவாகும்.

ல, ள, ழ

ர, ற

ண, ந, ன

7.5.2.2.5 அண்மை விசைப் பிழைகள்

தட்டச்சு செய்யும் போது பயன்படுத்துவோர் சரியான ஒன்றிற்குப் பதிலாகத் தவறான ஒன்றைத் தட்டச்சு செய்யலாம். எனவே தவறாகத் தட்டச்சு செய்யப்பட்ட எழுத்தின் அடுத்துவரும் விசைகளைக் கருத்தில் கொள்வது முக்கியமான பங்களிப்பு செய்கின்றது. தவறாகத் தட்டச்சு செய்யப்பட்ட எழுத்து ஏதாவது அண்மை விசையுடன் பொருந்தினால் தவறான எழுத்துக்குப் பதிலாகச் சரியானதை இடம்பெயர்க்கவும். அதன்பின்னர் திருத்தப்பட்ட சொல்லைச் சாத்தியமான பரிந்துரைகளை உருவாக்க அகராதியில் அல்லது வேற்றுமை விதிகளால் சரிபார்க்கவும். அண்மை விசை திருத்தங்கள் தமிழ் 99 என்ற பண்பட்ட விசைப்பலகை அடிப்படையில் அமையும்.

எடுத்துக்காட்டாக *'அவனுடைய'* என்ற சொல் *அவனுடஎய* அல்லது *அவனுடோய* என்று தவறாகத் தட்டச்சு செய்யப்படக்கூடும். இவைகள் *டை* என்ற எழுத்திற்குச் சாத்தியமான அண்மை விசைத் தவறுகளாகும். எல்லா சாத்தியமான அண்மை விசைகளும் இடம்பெயர்க்கப்பட்டு அகராதியில் சரிபார்க்கப்பட்டு தவறு திருத்தம் செயல்படுகின்றது.

எ.கா. அ -ஆ, ஏ, எ, அஒவ்

க -ள, ற, ப, ல ங

எ அ, ஆ, ஏ, உ, அஒவ், ஓ

ப ற, ந, க, ம, ல, ர

வேர்சொல்லில் உள்ள தவறுகளின் திருத்தம் முடிவுற்றால் தொடர்புள்ள ஒட்டுக்களுடன் திருத்தப்பட்டவைகள் பரிந்துரை உருவக்க நிலைக்கு அனுப்பப்படும்.

7.5.2.2.6 எழுத்து நிலைநிறுத்து விதிகள்

ஒரு தமிழ்ச் சொல் 12 உயிர்களில் ஒன்றாலோ 10 உயிர்மெய்களில் (க, ச, த, ந, ப, ம, ய, வ, ய, ஞ) ஒன்றாலோ தொடங்கப்படவேண்டும். இதுபோல் தமிழ்ச் சொல் 12 உயிர்களில் ஒன்றாலோ 9 மெய்களில் (ண், ம், ன், ய், ர், ல், ழ், ள், ன்) ஒன்றாலோ இறுதியுறவேண்டும். சொல் இந்த குழுமத்திலிருந்து தொடங்காவிட்டால், பிழைதிருத்தி சாத்தியமான திருத்தங்களைச் செய்யும்; அகராதியிலிருந்து அண்மையிலுள்ள பொருத்தமான வேர்சொற்களைப் பெறும். இது போல் ஒரு சொல் முன்னர் கூறின எழுத்துகளால் இறுதியுறாவிட்டால் வேற்றுமை விதிகளைப் பயன்படுத்தி பிழைதிருத்தி அதைத் திருத்தும்.

7.5.2.2.7அகராதியைப் பார்த்தல்

மற்றொரு முக்கியமான விசயம் அகராதியின் வடிவம், திரிபு மற்றும் ஆக்கத்தின் சிக்கல், அகராதி கோப்பு ஒழுங்கமைப்பு, அகராதி பிரிப்பு, சொல் பெறும் உபாயம்/தொழில்நுட்பம் மற்றும் பிறவற்றுடன் தொடர்புடைய கணினி அகராதி. இங்குப் பயன்படுத்தப்படும் தமிழ் அகராதி பெயர், வினை, பெயரடை, வினையடை, குறைச்சொற்கள் என்ற எல்லா வேர்ச்சொற்களையும் கொண்டிருக்கும். வேர்ச்சொல்லில் ஏதாவது பிழை இருந்தால் சொல்லின் திருத்தம் செய்யப்படும் மற்றும் அண்மைப்படும் பொருத்தமான அகராதி சொற்கள் சேகரிக்கப்படும். இந்த உருவாக்கம் பரிந்துரை உருவாக்க நிலைக்கு (suggestion generation phase) அனுப்பப்படும்.

7.5.2.2.8 பரிந்துரை உருவாக்கம்

பிழைதிருத்தி எல்லாச் சாத்தியமான பரிந்துரைகளையும் உருவாக்க உருபனியல் உருவாக்கியைப் பயன்படுத்தும். தமிழுக்குத் திட்டவரைவு செய்யப்பட்ட உருபனியல் உருவாக்கி பெயர்கள், வினைகள், பின்னுருபுகள், பெயரடைகள், வினையடைகள் மற்றும் பிற தொடரியல் வகைப்பாடுகளைத் தனித்தனியாகக் கையாள்வது அவசியாமாகும்; ஏனென்றால் இந்த ஒவ்வொரு தொடரியல் வகைப்பாட்டிற்கும் உருபனியல் கூறுகளின் சேர்க்கை வேறுபட்ட தகவல் வகை அடிப்படையிலானது.

வேர்களிலிருந்து பெயர்த் திரிபு ஆக்கங்களை உருவாக்கும் போது பன்மை ஒட்டின் வடிவைத் தீர்மானிக்கும் மொழியியல் விதிகள் கருத்தில் கொள்ளப்படவேண்டும். பெயர்களுக்கு வேற்றுமை உருபுகளை இணப்பது உருபனியல் உருவாக்கியின் முக்கியமான வேலையாகும். மேலும் பெயர்கள் வேற்றுமை உருபுகளைத் திரிபடியாகத்தான் ஏற்க இயலும் என்பதையும் கருத்தில் கொள்ளவேண்டும். ஒலிநிரவல் ஒட்டுகள் சில நேர்வுகளில் திரிபடிசார் ஒட்டுகளுடனோ (oblique suffix) பன்மை ஒட்டுகளுடனோ இணைந்து வரும். பெயர் வேர்களுடன் முன்னர் சொன்ன ஒட்டுகளின் சேர்க்கையின் போது புணர்ச்சி விதிகள் கருத்தில் கொள்ளப்படவேண்டும்.

தமிழ் வினையின் உருபனியல் அமைப்பு இடம், பால், எண் போன்றவற்றை உருப்பபடுத்தம் செய்யும் ஒட்டுகளைக் கையாளுவதாலும் வினையாற்றுவகை, வினைநோக்கு, காரணவாக்கம், வினைப்பாங்கு போன்றவற்றைக் காட்டும் துணைவினைகளுடன் இணைவதாலும் அதிக அளவில் கலவைநிலை அடைந்துள்ளது. வினையை உருபனியல் அடிப்படையில் உருவாக்கும் போது தெரிந்தெடுக்கப்பட்ட எழுவாயின் பால், எண், இடம் என்பனவற்றிகு ஏற்ப பொருத்தமான உடன்பாட்டு விகுதியைத் (வினைமுற்று விகுதியை) தெரிந்தெடுக்க வேண்டியது தேவையாகும்.

பிழை திருத்தியின் செயல்திறன் வழுநிலை பரிந்துரையாகத் தரப்படும் சரியான பரிந்துரை அடிப்படையில் அமையும். இந்நேர்வுகளில் பயன்படுத்துபவர் வழுநிலை பரிந்துரையை உறுதி செய்வது அவசியம்; பின்னர் அடுத்த தவறுக்கு முன்னேறலாம். இல்லையெனில் பயன்படுத்துபவர் பரிந்துரைகளின் பட்டியலைப் பார்த்து ஒன்றை தேர்ந்தெடுக்கவேண்டும். எல்லா பரிந்துரைகளும் அகராதியிலுள்ள சொற்களுடன் ஒப்பிடப்படவேண்டும். அகராதியிலுள்ள பரிந்துரைகள் பயன்படுத்துபவருக்குக் காட்டப்படும்.

7.5.2.3 முடிவுரை

தமிழுக்கான பிழைதிருத்தி தட்டச்சு செய்யும் போது ஏற்படும் பெருப்பாலான தவறுகளை அடையாளம் காண பயனாளிக்கு உதவுகின்றது. இந்தப் பிழை திருத்தியில் நடைமுறை செய்யப்பட்ட செயல்பாடுகள் பெயர்களின் வேற்றுமை உருபுகளையும் பின்னொட்டுகளையும் பிழை திருத்துதல், பெயர்களுக்குப் பெயரடைகளை திருத்துதல், வினைகளின் வினைமுற்று விகுதிகளைப் பிழைதிருத்துதல், வினையடை திருத்தம் மற்றும் அண்மை விசை பிழை திருத்தம் என்பனவாகும். தமிழ் பிழைதிருத்தியின் பயன்பாடு சொல் ஆய்விகள், தேடல் இயந்திரங்கள், தகவல் இறுப்பு மற்றும் தகவல் பிரித்தெடுப்பு ஒழுங்குமுறைகள், இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் என்பனவாகும்.

7.5.3 விஜய் சங்கர் ராமின் தமிழ் பிழைதிருத்தி உருவாக்குதல்

தமிழ் பிழை திருத்தி உருவக்குதல் குறித்து விஜய் சுந்தர் ராம் (Vija Sundar Ram 2012) தமது ஆய்வுக்கட்டுரையில் விளக்குகின்றார். பிழை திருத்தி பனுவல்களில் உள்ள தவறான சொற்களைக் கண்டுபிடிக்க உதவும் மற்றும் தவறான சொற்களுக்குச் சரியான மாற்றுச் சொற்களைத் தரும். முதல் செயல்பாடு “பிழை கண்டுபிடித்தல் (error detection)” எனப்படும்; இரண்டாவது செயல்பாடு “பிழைதிருத்தல் அல்லது பரிந்துரை உருவாக்குதல்” (error correction or suggestion generation) எனப்படும்.

7.5.3.1 எழுத்துப்பிழைதிருத்தலின் பலவேறு அணுகுமுறைகள்

விஜய் பின்வருவனவற்றை எழுத்துப்பிழை திருத்தலின் பல்வேறு அணுகுமுறைகளாகப் பட்டியலிடுகின்றார்.

- ஒலியல் அடிப்படையில் ஒன்றுபோல் இருக்கும் எழுத்துகளின் பிழைதிருத்தம்
- பிழையைக் கண்டுபிடித்தல் மற்றும் இருப்பிடம் கண்டுபிடித்தல்
- ஓசை ஊடக மாதிரி (noise channel model)
- ஒட்டுநிலை மொழிகளின் எழுத்துப்பிழை திருத்தம் (ஒஃபல்ஸர், கே 1994)
- எழுத்துப்பிழை திருத்த அணுகுமுறைக்குச் சாளர அடிப்படை அணுகுமுறை (Window based approach)
- எழுத்துப்பிழை திருத்தத்தில் நிகழ்தகவு மதிப்பெண்

7.5.3.2 இந்திய மொழிகளுக்கு இருக்கின்ற எழுத்துப் பிழை திருத்திகள்

பின்வரும் அட்டவணையில் இந்திய மொழிகளுக்கு இருக்கின்ற எழுத்துப் பிழை திருத்திகளை விஜய் பட்டியலிடுகின்றார்.

மொழிகள்	உருவாக்கப்பட்ட இடம்	பயன்படுத்தப்பட்ட அணுகுமுறை
அசாம் மொழி	RICILTS, கௌகாத்தி	அகராதி அடிப்படை அணுகுமுறை
வங்காள மொழி	இந்தியப் புள்ளியியல் நிறுவனம், கொல்கொத்தா (பி.பி. சௌத்ரி)	
தெலுங்கு கன்னட மொழிகள்	தெலுங்கு மூலவள மையம், ஹைதராபாத் பல்கலைக் கழகம்	உருபனியல் ஆய்வி அடிப்படை அணுகுமுறை
மலையாளம்	RICILTS, திருவனந்தபுரம்	விதி மற்றும் அகராதி அடிப்படை அணுகுமுறை
மராத்தி	RICILTS, ஐஐடி பாம்பே	அகராதி மற்றும் பைகிராம் மற்றும் டிரைகிராம் அணுகுமுறை
தமிழ்	தனபாலன், அண்ணா பல்கலைக் கழகம், எயுகேபிசி, சென்னைப் பல்கலைக் கழகம்	உருபனியல் பகுப்பாய்வியைப் பயன்படுத்தி தரவுத்தொகுதிஅடிப்படை + உருபனியல் பகுப்பாய்வி விதிகள்

7.5.3.3 பிழை கண்டுபிடித்தல்

இதை இருவழிகளில் செய்ய இயலும்

1. சொல்லை அகராதியில் ஒப்பிட்டுப் பார்த்தல் (அகராதி அணுகுமுறை)
2. சொல்லை உருபனியல் ஆய்வியால் ஆய்ந்து பிழை கண்டுபிடித்தல் (மொழியியல் அணுகுமுறை)

இரண்டு அணுகுமுறைகள் உள்ளன :

1. தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை (முற்றுநிலைத் தானியங்கியைப் பயன்படுத்தி)
2. மொழியியல் அடிப்படையிலான அணுகுமுறை (உருபனியல் ஆய்வியைப் பயன்படுத்தி;

7.5.3.3.1 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை

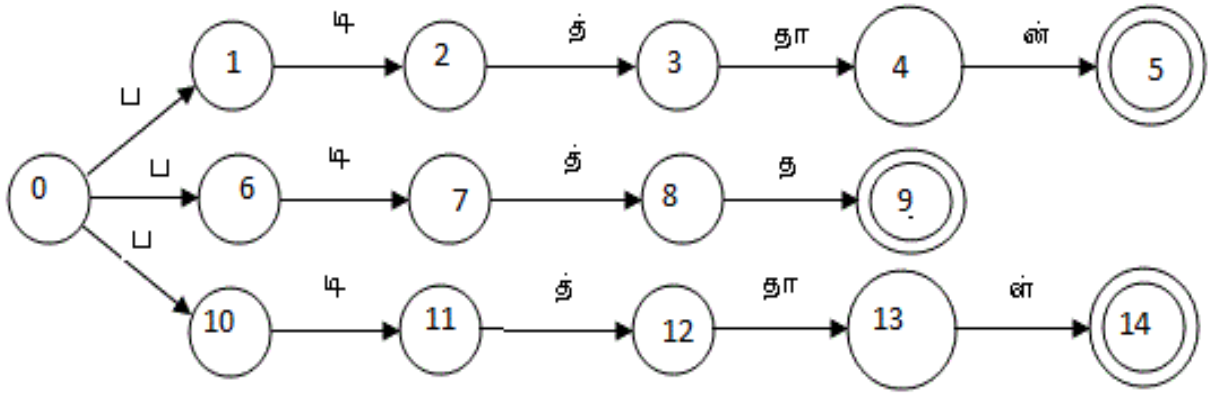
தரவுத்தொகுதி முற்றுநிலைத் தானியங்கியால் (Finite State Automata) உருப்படுத்தம் செய்யப்படும். இது விரைவக மதிப்பீடு செய்யும். ஒரு கச்சிதமான உருப்படுத்தத்தைச் செய்யும். சொல்லின் ஒவ்வொரு எழுத்தும் முற்றுநிலைத் தானியியங்கியில் நிலைமாற்றக் குறியீடாக (transition symbol) இருக்கும்.

முற்றுநிலைத் தானியங்கி

முற்றுநிலைத் தானியங்கி பனுவல் கோர்வையில் எளிதான தொடரியல் அமைப்புகளை அல்லது ஒழுங்கு அமைப்புகளை புரிந்துகொள்வதற்காகப் பயன்படுத்தப்படும் அருவ உபாயம் ஆகும். ஒரு தனியிங்கி நிலை படம் (state diagram) என்று அழைக்கப்படுகின்ற வழிகாட்டப்பட்ட வரைபடத்தால் (directed graph) வழக்கமாக விவரிக்கப்படும். முற்றுநிலைத் தானியங்கி ஒரு கோர்வை ஆய்வு உபாயம் (string processing device) ஆகும்; அது ஒரு பனுவல் கோர்வையை உள்ளீடாக ஏற்கும் மற்றும் கோர்வையை ஏற்கவோ மறுக்கவோ செய்யும். கணித அடிப்படையில் இதை ஒரு குழுமத்திற்கு {ஏல், மறு} ஒரு குழுமக் கோர்வைகளைப் பொருத்தும் ஒரு செயல்பாடாகக் கருதலாம்.

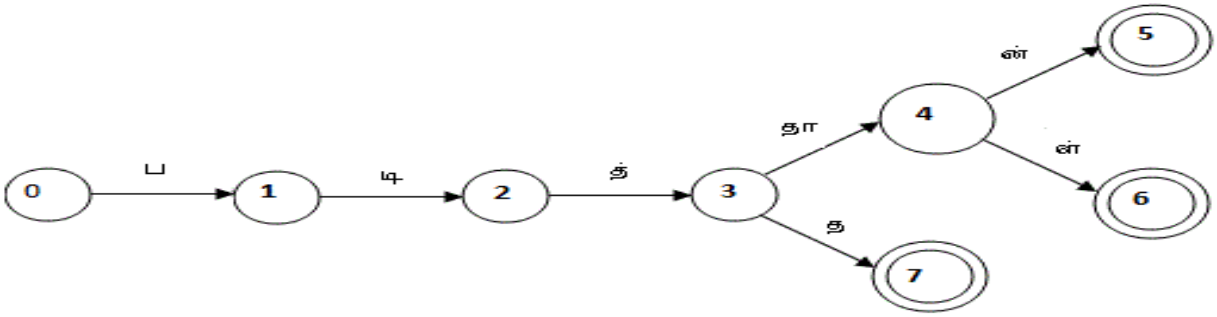
நிலைகள் வட்டங்களால் உருப்படுத்தம் செய்யப்படும். உள்ளீடாக ஒரு குறியீடு உபாயத்தால் பெறப்பட்டால் நேர்கின்ற நிலைமாற்றங்களை (transitions) அம்புக்குறிகள் உருப்படுத்தம் செய்யும். தொடக்கத்தில் தீர்மானிக்கப்படாத (non-diterministic) முற்றுநிலைத் தானியங்கி உருவாக்கப்பட்டது; இது பின்னர் முற்றுநிலைத் தானியங்கி கருவித்தொகுதியைப் (tool kit) பயன்படுத்தித் தீர்மானிக்கப்பட்ட (diterministic) மற்றும் குறுக்கப்பட்ட (minimized) முற்றுநிலைத் தானியங்கியாக ஆக்கப்பட்டது.

தீர்மானிக்கப்படாத முற்றுநிலைத்தானியங்கி



ஒரே குறியீடுகள் பல நிலைகளுக்குத் தூண்ட இயலும். பல வழிகள் சாத்தியமாகும். கணினிச் செயல்பாடுகளுக்கு பொருந்தாது.

தீர்மானிக்கப்பட்ட முற்றுநிலைத் தானியங்கி



ஒரு குறியீடு ஒரு நிலையிலிருந்து மற்றொரு நிலைக்கு மட்டுமே தூண்ட இயலும். பல வழிகள் இல்லை. இதன் அனுகூலம் முனைப்பான உருப்படுத்தமும் விரைவான செயற்பாங்கும் ஆகும்.

நிலை அட்டவணை உருப்படுத்தம்

ஆரம்பநிலை	**அடுத்த நிலை**	**நிலைமாற்றக் குறியீடு**
0	1	ப
1	2	டி
2	3	த்
3	4	தா
3	7	த
7	இறுதிநிலை	

4	5	ன்
4	6	ள்
5	இறுதிநிலை	
6	இறுதிநிலை	

7.5.3.3.2. மொழியியல் அடிப்படையிலான அணுகுமுறை

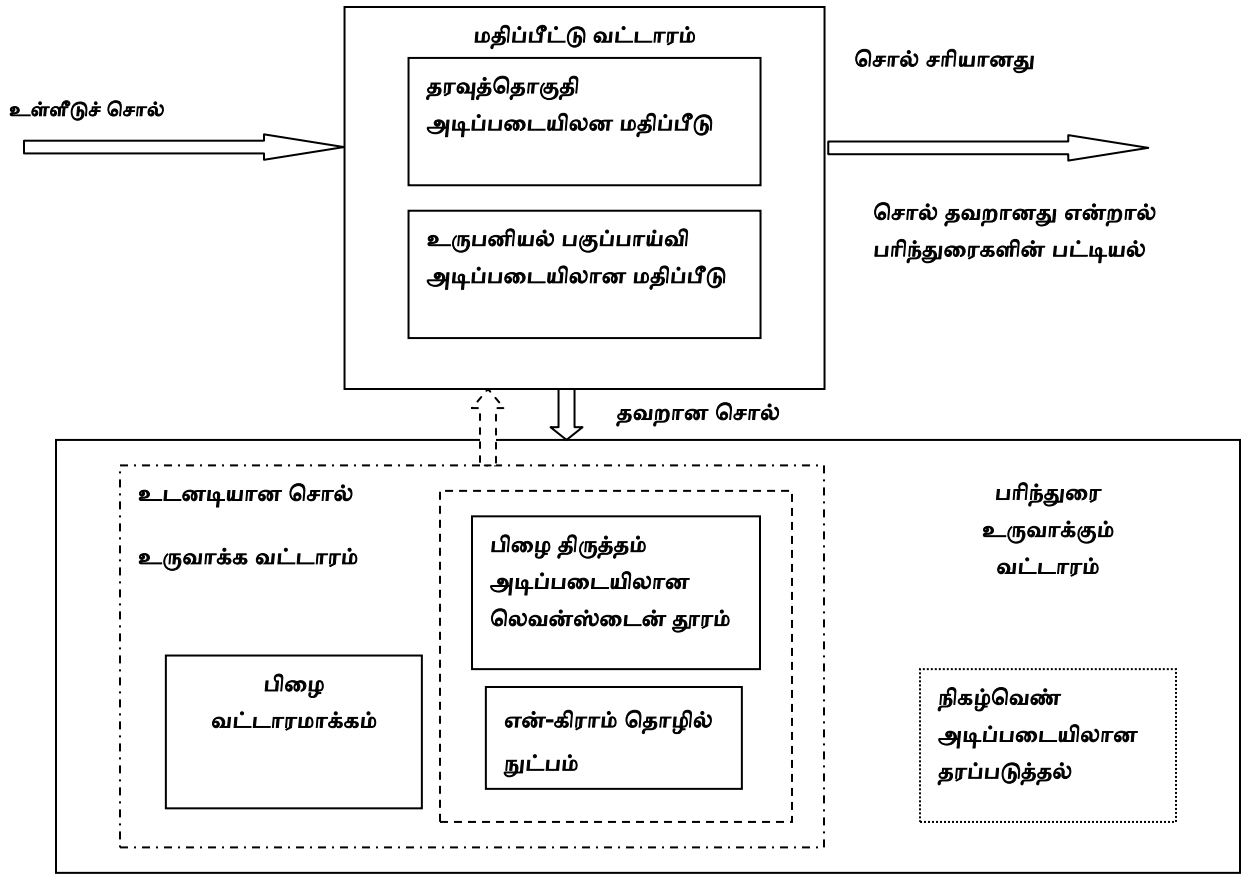
உருபனியல் பகுப்பாய்வி

உருபனியல் பகுப்பாய்வு என்பது சொல்லை அதில் உள்ளடங்கும் உருபன்களாகப் பிரிதல் ஆகும். இச்செயலில் உருபன்களின் அடையாளங்கள் தேவையில்லை. சொல்லுக்கு ஏற்கத்தக்க பகுப்பாய்வு இருந்தால் அது ஏற்கத்தக்கச் சொல்லாகக் கருதப்படும்.

படித்தான் = படி + த்த் + ஆன்

பின்னொட்டு நீக்கி (suffix stripper) இச்செயன்மைக்குப் போதுமானதா?

பிழை திருத்தியின் கட்டமைப்பு



பிழை திருத்தல்

பிழை திருத்துவதற்குப் சவுண்டெக்ஸ் கொள்கை (soundex principle) போன்ற பல அணுகுமுறைகள் பின்பற்றப் படுகின்றன. சொல் தவறு இன்று கண்டுபிடிக்கப்பட்டால் பின்வருவனவற்றைப் பயன்படுத்தி பரித்துரைகள் உருவாக்கப்படுகின்றன.

- பிழையை வட்டார எல்லைக்கு உட்படுத்தல்
- திருத்த தூரத்தையும் என்கிராமையும் பயன்படுத்தி இடைப்பட்ட சொற்களை உருவாக்குதல்
- அதை மதிப்பிடுதல்
- தரப்படுத்துதல்

பிழை திருத்தம்

இது பின்வரும் தொழில் நுட்பங்களை உள்ளடக்கும்

1. லெவன்ஸ்டைன் தூரம்
2. பிழையை எல்லைக்குட்படுத்தல்
3. என்-கிராம் தொழில் நுட்பம்
4. நிகழ்வெண் அடிப்படையிலான தரப்படுத்தும் படிமுறை

லெவன்ஸ்டைன் தூரம்

லெவன்ஸ்டைன் தூரம் (Levenstein distance) என்பது வலாடிமிர் லெவன்ஸ்டைன் என்ற ரஷ்ய அறிவியல் அறிஞரின் நினைவாகப் பெயரிடப்பட்டுள்ளது. பதின்ம அளவுமுறையும் சிலவேளைகளில் திருத்து தூரம் (edit distance) என்று அழைக்கப்படும். லெவன்ஸ்டைன் தூரம் (LD) என்பது இரு கோர்வைகளுக்கு இடையிலுள்ள ஒற்றுமை ஆகும்; இதில் நாம் மூலக் கோர்வையை s என்றும் இலக்குக் கோர்வையை t என்றும் குறிப்பிடுகின்றோம்.

t ஆக மாற்றத் தேவைப்படும் நீக்கல் (deletion), செருகல் (insertion) அல்லது பதிலீடுசெய்தல் (substitution) என்ற நடவடிக்கையின் எண்ணிக்கையின் வடிவில் தூரம் வெளிப்படுத்தப்படுகின்றது. எடுத்துக்காட்டாக, s என்பது “test” மற்றும் t என்பது “test” என்றால் LD (s,t) = 0, ஏனென்றால் எந்த மாற்றமும் தேவையில்லை. இரண்டு கோர்வைகளும் ஒன்றாக இருக்கின்றன. s என்பது “test” மற்றும் t என்பது “tent” என்றால் LD (s,t) = 1, ஏனென்றால் s-ஐ t-ஆக மாற்ற ஒரு பதிலீடு முறை (“s”-ஐ “n”–ஆக மாற்ற) போதுமானது.

மூலச் சொல்: படித்த

இலக்குச்சொல்: படித்தான்

லெவன்ஸ்டைன் தூரம் 2

மூலச் சொல்: படித்தான்

இலக்குச் சொல்: படித்தாள்

லெவன்ஸ்டைன் தூரம் 1

பிழை எல்லைப்படுத்தல்

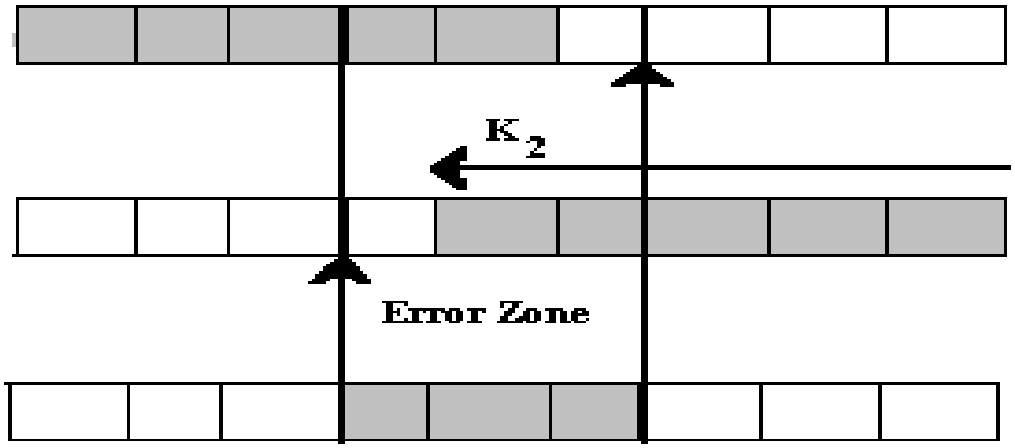
பிழை செய்யப்பட்ட சொல்லின் பிழை பரப்பைக் (error zone) கண்டுபிடிக்கும் நுட்பம் பிழை எல்லைப்படுத்தல் எனப்படும். கட்டுப்பாட்டிற்குட்பட்ட செயற்பாங்கை நிறைவேற்றவும் எதிர்பார்க்கப்பட்ட மாற்றுப் பரிந்துரைகளை உருவாக்கவும் பிழை எல்லைப்படுத்தம் தேவையாகும்.

நாம் நமது செயல்பாட்டில் பின்வரும் ஊகங்களைச் செய்கின்றோம்:

1. சொல்லில் செருகல், நீக்கல், பதிலீடு செய்தல், இடம் மாறுதல் என்பனவற்றுள் ஒரு தவறு தான் இருக்கக்கூடும்.
2. சரியான சொல் இரு அகராதிகளிலும் (மரபு மற்றும் திருப்பப்பட்ட சொல்) கோப்புகளிலும் இருக்கும்.

D_c என்ற மரபு அகராதியும் R_c என்ற மாற்றப்பட்ட அகராதியும் பயன்படுத்தப்படுகின்றன. $W = x_1x_2..x_k$ என்ற சொல்லின் மாற்றப்பட்ட சொல் $W_r = x_kx_{k-1}…x_2x_1$ ஆகும், இதில் x_1 என்பது

ஏற்கத்தக்க தமிழ் எழுத்தாகும். R_c என்ற மாற்றப்பட்ட அகராதியில் மரபு அகராதி D_c-யின் எல்லாச் சொற்களின் திருப்பப்பட்ட பதிப்புருவும் வைத்திருக்கப்படும்.



மரபு மற்றும் மாற்ற அகராதி அடிப்படையில் பிழை எல்லைப்படுதல்

இது மூன்று விதமாகச் செய்யப்படுகின்றது.

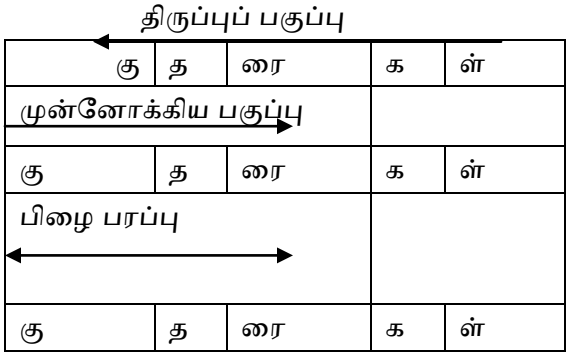
(அ) மரபு அகராதி மூலம் பிழை எல்லைப்படுத்தல்

(ஆ) திருப்பு அகராதி மூலம் பிழை எல்லைப்படுத்தம்

(இ) இரு அகராதியும் பயன்படுத்தப்படுகையில் பிழை எல்லைப்படுத்தம்

சரியான சொல்: குதிரைகள்

தவறான உள்ளீட்டுச் சொல்: குதரகள்



தவறு பரப்பில் திருத்தம்

இது என்-கிராம் நுட்பத்தைப் பயன்படுத்திச் செய்யப்படுகின்றது. என்-கிராம் இலக்கணம் என்-ஆவது நிரல் மார்க்கொவ் மொழி மாதிரியின் உருப்படுத்தம் ஆகும்; இதில் ஒரு குறியீட்டின் நேர்வு சாத்தியம் பிற குறியீட்டின் N -1-ஆவதன் முந்தைய நேர்வால் கட்டுப்படுத்தப்பட்டுள்ளது. ஒரு W என்ற தரப்பட்ட சொல் வரிசையின் காரணகாரிய சாத்தியத்துடன் P(W) புரிந்துகொள்வானைத் தருவதற்கு என்-கிராம் மாதிரிகள் பெரிய சொற்றொகை பேச்சு

புரிந்துகொள்ளல் ஒழுங்கமைப்புகளில் பயன்படுத்தப்படுகின்றது. என்–கிராம் மொழி மாதிரிகள் எதிர்பார்க்கப்பட்ட உள்ளீடாக மொழியின் ஒரே பண்புகளைப் பங்கிட்டுக்கொள்ளும் பெரிய பயிற்சி பனுவல்களிலிருந்து வழக்கமாக ஆக்கப்படுகின்றது. இந்தத் தகவல் பேசுபவர்களின் பேசு உருவாக்கிகளின் பண்புக்கூறுகளை மாதிரிசெய்யும் ஒலியியக்கவியல் மாதிரி P(W|O) என்பதை குறைநிறவாக்கம் செய்யும். இக்கூறுகள் இரண்டும் இணைந்து W' = armax$_w$ P(W|O) என்ற பெரும்பான்மை சாத்தியமான உள்ளீட்டு வரிசையை கணிக்க ஒரு ஒழுங்குமுறையை அனுமதிக்கும்; இதில் W' = armax$_w$ P(W|O) P (W) ஆக O என்பது உள்ளீட்டு சமிக்கையாகும். நாம் என்-கிராமை சொல்நிலையில் அல்லாமல் எழுத்துநிலையில் பயன்படுத்துகின்றோம்; நாம் N= 1,2,3 என்று எடுத்துக்கொள்கின்றோம். பைகிராம் நுட்பத்தில் முதலில் வரும் எழுதுக்கு எல்லாச் சாத்தியமான அடுத்த எழுத்தைக் கண்டுபிடிக்க முயலுகின்றோம். டிரைகிராம் நுட்பத்தில் ஒரு குறிப்பிட்ட முதல் மற்றும் மூன்றாவது எழுத்துக்குச் சாத்தியமான எல்லா மைய எழுத்தை கண்டுபிடிக்க முயலுகின்றோம்.

ஆக்கப்பட்ட சொற்களுக்களின் மதிப்பீடு

பிழை எல்லைப்படுத்தல் நுட்பங்களைப் பயன்படுத்தி பிழை பரப்புகள் அடையாளம் காணப்படும். என் கிராம் நுட்பத்தைப் பயன்படுத்தி இடைநிலைச் சொற்கள் உருவாக்கப்படும். இந்த இடைநிலைச் சொற்கள் பின்வருவனவற்றைப் பயன்படுத்தி மதிப்பீடு செய்யப்படும்:

- முற்றுநிலைத் தானியங்கி அடிப்படையிலான தரவுத்தொகுதி அணுகுமுறை
- உருபனியல் பகுப்பாய்வி

உருவாக்கப்பட்ட சொற்களின் முன்னிலைப்படுத்தல்

உருவாக்கப்பட்ட சொற்கள் தரப்படுத்தப்படவேண்டும். தரப்படுத்தம் பின்வரும் மூன்று அடிப்படையில் அமையும்:

- நிகழ்வெண் அடிப்படையில்
- பயன்பாட்டாளர் அடிப்படையில்
- பொரு`ண்மைக்களம் அடிப்படையில்

பிழை திருத்தியின் பயன்பாடு

பனுவல் செம்மையாக்கிகளில் பயன்படுத்தப்படுகின்றது. தேடல் இயந்திரங்களிலும் போர்டல்களிலும் பயன்படுத்தப்படுகின்றது

7.5.4 தெய்வசுந்தரத்தின் மென் தமிழ் சொல்லாளர்

தெய்வசுந்தரம் தமது மொழியியல் அறிவையும் கணிப்பொறி மொழியியல் அறிவையும் பயன்படுத்தி பல்லாண்டு முயற்சிகளுக்குப் பின் மிக சிறப்பான ஒரு தமிழ் சொல்லாய்வியை உருவாக்கியுள்ளார். இது சொற்பிழை திருத்தி, சந்திப்பிழை திருத்தி, மொழிமாற்றி என்ற பகுதிகளைக் கொண்டது. இது இயல்மொழி ஆய்வு, கணினி மொழியியல் ஆய்வு, மொழித்தொழில் நுட்பம் ஆகியவற்றின் அடிப்படையில் உருவாக்கப்பட்ட ஒரு மென்மம் ஆகும்.

இதில் தமிழை உள்ளீடு செய்ய பலவகையான விசைப்பலகைகள் உள்ளன. ஒருங்குறி எழுத்துருக்களைப் பயன்படுத்தும் வசதி உள்ளது. எந்த வகைக் குறியீட்டையும் வடிவமைப்பு மாறாமல் எந்த குறியீட்டுக்கும் மாற்றலாம். அகராதிச் சொற்கள் மட்டுமல்லாமல் விகுதிகள் ஏற்ற சொற்களையும் ஆங்கிலச் சொற்களையும் திருத்தித்தரும். ஒற்று (சந்தி) இடாத இடங்களில் இடும். தேவையில்லாத இடங்களில் ஒற்றினை நீக்கியும் தரும். ஆங்கிலச் சொற்களை ஒலிபெயர்த்து எழுதினாலும் அவற்றுக்கு நிகரான தமிழ்ச் சொற்களாக மாறித்தரும். ஆறுவகையான அகரதிகள் உள்ளன: 45 ஆயிரம் தற்காலத் தமிழ்ச் சொற்களைக் கொண்ட தமிழ்-ஆங்கிலம், ஆங்கிலம்-தமிழகராதி, இணைச் சொல் அகராதி (இது ஒரு சொல்லுக்கு இணையான பல சொற்களை வழங்கும்), எதிர்சொல் அகராதி (இது ஒரு சொல்லுக்குரிய எதிர்ச்சொற்களை வழங்கும்), மயங்கொலிச்சொல் அகராதி (தமிழில் ஏற்படும் *ல, ள, ழ* மயக்கங்கள், *ந, ண, ன* மயக்கங்கள், *ர, ற* மயக்கங்கள் ஆகியவற்றைத் தவிர்க்கப் பயன்படும் பொருள் வேறுபாட்டுக் குறிப்புகள் அடங்கியது), தமிழ்நாட்டு அரசு ஆட்சிச்சொல் அகராதி. மேலும் அகரவரிசைப்படுத்தி, சொல்லடைவு, எண்-எழுத்து மாற்றி (பயன்பாட்டில் இருக்கும் எத்தகைய எண்களையும் தமிப் புணர்ச்சி விதிப்படி தமிழ்சொற்களாக மாற்றித்தரும், கோப்புமாற்றி போன்ற பலவித வசதிகளைக் கொண்டு இயங்குவது.

இது முற்றிலும் மொழியியல் கோட்பாடுகளைப் பயன்படுத்தி இரண்டு பதின்ம ஆண்டுகளின் கடின உழைப்பால் உருவாக்கப்பட்ட சொல்லாளராகும். இது தமிழக அரசால் ஏற்றுக்கொள்ளப்பட்டு தமிழ்நாட்டில் அலுவலங்களிலும் பிற இடங்களிலும் பயன்படுத்தப்படும்படி தமிழக அரசால் ஊக்குவிக்கப்பட்டுள்ளது.

7.5.5. எடுத்துக்காட்டான பிழை திருத்தி

இங்கு ஒரு எடுத்துக்காட்டான எழுத்து மற்றும் இலக்கணத் திருத்தி உருவாக்குவது குறித்த தொழில் நுட்பம் கூறப்பட்டுள்ளது. இது தரவுத்தொகுதி அடிப்படையிலான பிழைதிருத்தி. தரப்பட்டுள்ள தரவுத்தொகுதி அடிப்படையில் இது இயங்கும்.

பிழைதிருத்தியின் வழிமுறை அமைப்பு

சொல் தரப்படுகையில் நாம் அச்சொல்லின் சாத்தியமான எழுத்துப் பிழை திருத்தத்தைத் தேர்ந்தெடுக்க முயலுவோம். (அந்தச் சரியான சொல் மூலச்சொல்லாகவே இருக்கலாம்.) இதை உறுதியாக அறிவதற்கு வழியில்லை; எனவே நாம் சாத்தியங்களைப் பரிந்துரைக்கின்றோம். எடுத்துக்காட்டாக சாத்தியமான எல்லா திருத்தங்களிலிருந்தும் சரியான c என்ற திருத்தத்தைக் கண்டுபிடிப்பதற்கு முயலுவோம்; இது மூலச்சொல் தரப்படுகையில் c-யின் சாத்தியங்களை மேம்படுத்தும்:

argmaxc P(c|w)

By Bayes' Theorem this is equivalent to:

argmaxc P(w|c) P(c) / P(w)

Since P(w) is the same for every possible c, we can ignore it, giving:

argmaxc P(w|c) P(c)

இந்த வெளிப்படுத்தத்தின் மூன்று பகுதிகள் உள்ளன. வலமிருந்து இடமாக நமக்குப் பின்வருவது இருக்கின்றது:

1.P (c), பரிந்துரைக்கப்பட்ட திருத்தம் c தானாகவே நிலைத்துநிற்கும் என்ற சாத்தியம். இது மொழி மாதிரி என்று அழைக்கப்படும்: "தமிழ்ப் பனுவலில் c தோன்றுவது எந்த அளவுக்குச் சாத்தியம்?" என்ற கேள்விக்கு விடை அளிப்பதாக எண்ணிக்கொள்ளவும்; P("zxzxzxzyyy") என்பது பூஜியத்திற்கு அண்மைப்படும்.

2. P(w|c), ஒருவர் c-ஐ குறிப்பிட எண்ணும் போது w என்பது தட்டச்சு செய்யப்படும் சாத்தியம். இது பிழை மாதிரியாகும்: "ஒருவர் c-ஐ குறிப்பிட விரும்புகையில் தவறாக w-ஐத் தட்டச்சு செய்யவது எந்த அளவுக்குச் சாத்தியமானது?" என்ற கேள்விக்கு விடையளிப்பதாக எண்ணிக்கொள்ளவும்.

3. கட்டுப்பாடு உபாயம், argmaxc, c-யின் எல்லாச் சாத்தியமான மதிப்புகளைப் பட்டியலிடவும் மற்றும் மிகச் சரியான ஒருங்கிணைந்த சாத்திய மதிப்பெண்ணை தெரிந்தெடுக்கவும் கூறும் கட்டுப்பாட்டு உபாயம்.

வெளிப்படையான கேள்வி பின்வருவதாகும்: நாம் P(c/w) என்ற எளிய வெளிப்படுத்தத்தை எடுத்துக்கொண்டு ஒரு மாதிரிக்குப் பதிலாக இரு மாதிரிகளை உட்படுத்தும் கூடுதல் கலவைத் தன்மையான வெளிப்பாட்டால் இடம்பெயர்க்கின்றோம்? இதன் விடை (P(c/w) ஏற்கனவே இரு காரணிகளை ஒருங்கிணைக்கின்றது மற்றும் இரண்டையும் பிரிப்பதும் அவற்றை வெளிப்படையாகக் கையாளுவதும் எளிது. w= "பட்தம்" என்ற தவறான சொல்லைக் கருத்தில் கொள்ளவும்; இதற்கு குறைந்தது இரண்டு திருத்தங்கள் இருக்கலாம்: c= படம், c = பட்டம். இவற்றில் எதற்குக் கூடுதல் P(c/w) மதிப்பு இருக்கின்றது? "த" என்பதை "ட" என மாற்றினால் போதுமானது என்ற சிறிய மாற்றம் என்ற அளவில் "பட்டம்" சரியானதாகத் தோன்றுகின்றது. மாறாக, "படம்" என்பது மிகப் பொதுவாகப் பயன்படுத்தப்படுகின்ற சொல் என்ற நிலையிலும் விசைப் அண்மை என்ற நிலையிலும் படம் என்பதைத் தவறாக தட்டச்சு செய்ய வாய்ப்பு இருக்கின்றது. முக்கியமாத விசயம் நாம் c என்பதன் சாத்தியத்தையும் c என்பது w என மாறுவதன் சாத்தியத்தையும் கருத்தில் கொள்ளவேண்டும்; எனவே இவ்விரு காரணிகளைப் முறையாகப் பிரிப்பது தெளிவானது. இபோது நாம் எவ்வாறு வழிமுறை வரைவு வேலை செய்கின்றது என்பதைக் காட்டுவதற்குத் தயாராய் உள்ளோம்.

நாம் ஒரு பெரிய பனுவல் கோப்பைப் (big.txt) படிப்போம்; இது ஒரு மில்லியன் சொற்களைக் கொண்டிருக்கும். இது பல வித தரவுகள் சேர்ந்த தொகுதியாகும்.

பின்னர் நாம் கோப்பிலிருந்து தனிச் சொற்களைப் பிரித்தெடுப்போம் (இதன் படி 'படம்', 'பட்டம்' என்ற சொற்கள் ஒன்றாகும் மற்றும் சொற்கள் எழுத்துக்களின் கோர்வையாகும்.) அடுத்தபடியாக சாத்திய மாதிரியைப் பயிற்சிசெய்யவும்; இது செயல்பாட்டுப் பயிற்சியைப் பயன்படுத்தி எவ்வளவு தடவை ஒவ்வொரு சொல்லும் வருகின்றது என்பதைக் கூறுவதாகும். பின்வருதைப் பார்க்கவும்:

```
def words(text): return re.findall('[a-z]+', text.lower())

def train(features):
    model = collections.defaultdict(lambda: 1)
   for f in features:
  model[f] += 1
 return model

NWORDS = train(words(file('big.txt').read()))
```

இந்த இடத்தில் NWORDS[w] என்பது w எத்தனை தடவை காணப்படுகின்றது என்ற கணக்கைக் கொண்டிருக்கின்றது. ஒரு தொகுப்பு இருக்கின்றது: புதிய சொற்கள். தமிழில் நல்ல சொல் பயிற்சித் தரவில் இல்லாவிடில் என்ன நடக்கும் என்பது கோள்வி. இச்சிக்கலுக்குப் பல தரமான அணுகுமுறைகள் உள்ளன; நாம் எளிதான ஒன்றை எடுக்கின்றோம்; இது புதிய சொற்களை நாம் அவற்றை ஒருதடவைப் பார்த்திருப்தாகக் கையாளும். இந்த பொதுவான செயல்முறை சீராக்கம் (smoothing) எனப்படும்; ஏனென்றால் நாம் பூஜியமாக இருக்க இயலும் சாத்திய வினியோகத்தின் பகுதிகளை மிகக் குறைவான சாத்திய எண்ணிக்கைக்கு மாறும் படிக்கு சீர்மை செய்வோம். இது வகுப்புச் சேகரிப்புகள், defaultdict வழியாக எய்தப்படுத்துகின்றது; இது பைத்தான் dict (பிற மொழிகள் இதை hash table என்று அழைக்கும்) போன்று சீரானது; நாம் எந்த விசையின் வழுநிலை மதிப்பையும் குறிப்பிட இயலும் என்பது விதிவிலக்காகும்; இங்கு நாம் 1-ஐப் பயன்படுத்துகின்றோம்.

இப்போது நாம் w என்ற சொல்லின் சாத்தியமான திருத்தங்கள் c-ஐப் பட்டியலிடும் சிக்கலைப் பார்ப்போம். இரு சொற்களுக்கு இடையிலுள்ள திருத்து தூரத்தை (edit distance) பற்றி பேசுவது பொதுவானதாகும்: ஒன்றை மற்றொன்றாக மாற்றுவதற்கு அது எடுக்கும் திருத்தங்களின் எண்ணிக்கை. நிக்குவது (ஒரு எழுத்தை நீக்குவது), இடம்மாற்றுவது (அண்மையில் வரும் எழுத்துக்களை இடம் மாற்றுவது), மாற்றுவது (ஒரு எழுத்தை மற்றொரு எழுத்தாக மாற்றுவது) என திருத்தம் அமையலாம். கீழே வருவது w என்பதிலிருந்து ஒரு திருத்த தூரத்தில் உள்ள எல்லா c சொற்களின் குழுமத்தை திருப்பி அனுப்பும் செயல்பாடாகும்.

```
def edits1(word):
splits     = [(word[:i], word[i:]) for i in range(len(word) + 1)]
deletes    = [a + b[1:] for a, b in splits if b]
transposes = [a + b[1] + b[0] + b[2:] for a, b in splits if len(b)>1]
replaces   = [a + c + b[1:] for a, b in splits for c in alphabet if b]
inserts    = [a + c + b     for a, b in splits for c in alphabet]
return set(deletes + transposes + replaces + inserts)
```

இது ஒரு பெரிய குழுமம் ஆகும். n நீளம் உள்ள ஒரு சொல்லுக்கு n நீக்கங்கள், n-1 இடமாற்றங்கள், 26n மாற்றங்கள், 26(n+1) செருகல்கள் இருக்கும்; மொத்தம் 54n+25 ஆகும் (இவற்றில் சில இரட்டிப்புகளாகும். எடுத்துக்காட்டாக, len(edits1 ('something'))--, அதாவது edits1('something') என்பதன் தனிமங்களின் எண்ணிக்கை 494 ஆகும்.

பிழை திருத்தம் பற்றிய இலக்கியங்கள் எழுத்துப் பிழைகளின் 80%-இலிருந்து 90% வரை இலக்கிலிருந்து 1 திருத்து தூரத்தில் இருக்கின்றது என்று கூறுகின்றன. நாம் 2 திருத்து தூரத்தைக் கருத்தில் கொள்ளவது தேவையாகும். அது எளிதாகும்: எல்லா edits1-இன் விளைவுகளுக்கும் edits1-ஐப் பிரயோகிக்கவும்.

```
def edits2(word):
return set(e2 for e1 in edits1(word) for e2 in edits1(e1))
```

இது எழுதுவதற்கு எளிதாகும், ஆனால் நாம் சில ஆபத்தான சிக்கல்களில் சிக்கிக்கொள்வதற்குத் தொடங்குகின்றோம்: len(edits2('something')) is 114,324. இருப்பினும் நமக்கு நல்ல செயலெல்லை கிடைக்கின்றது; 270 பரிசோதனை நேர்வுகளில் 3 தான் 2-ஐவிட கூடுதல் திருத்து தூரத்தைக் கொண்டுள்ளன. அதாவது edits2 நேர்வுகளின் 98%-ஐ உட்படுத்துகின்றது; அதாவது அது போதுமானதாகும். நாம் திருத்து தூரம் 2-ஐத் தாண்டி செல்லமாட்டோம் என்பதால் நாம் சிறிய உகப்பாக்கத்தைச் செய்ய இயலும்: சொற்கள் என்று உண்மையில் கூறப்படும் கருதத்தக்கவைகளை (candidates) மட்டும் வைத்திருக்கவும்; நாம் இன்னும் எல்லாச் சாத்தியங்களையும் கருத்தில் கொள்ளவேண்டும்; ஆனால் நாம் அவற்றின் பெரிய குழுமத்தைக் கட்டத் தேவை இல்லை. known_edits2 என்ற செயல்பாடு பின்வருவதைச் செய்யும்:

```
def known_edits2(word):
return set(e2 for e1 in edits1(word) for e2 in edits1(e1) if e2 in NWORDS)
```

எடுத்துக்காட்டாக இப்போது known_edits2(‘something’) 4 சொற்களின் குழுமமாகும்: edits2-ஆல் உருவாக்கப்பட்ட 114,324 சொற்களின் குழுமத்திற்குப் பதிலாக {'smoothing', 'seething', 'something', 'soothing'}ஆகும். இது காரியங்களை கிட்டத்தட்ட 10% வரை விரைவுபடுத்தும்.

இப்பொழுது விடுபட்ட ஒரே ஒரு பகுதி பிழை மாதிரி P(w/c) ஆகும். இங்கு நாம் சிக்கலில் அகப்படுவோம். திருத்து தூரம் 1-இன் எல்லா தெரிந்த சொற்களும் திருத்து தூரம் 2-இன்

தெரிந்த சொற்களை விட முடிவற்ற கூடுதல் சாத்தியமாகும். மற்றும் திருத்து தூரம் 0-இன் தெரிந்த சொற்களை விட குறைத்த சாத்தியமாகும். 'தெரிந்த சொல்' என்பது மொழி மாதிரி பயிற்சி தரவில் நாம் கண்டிருப்பவை -- அகராதியில் உள்ள ஒரு சொல். நாம் இந்த நடவடிக்கையைப் பின்வருமாரு செயல்படுத்தலாம்:

```
def known(words): return set(w for w in words if w in NWORDS)
def correct(word):
candidates = known([word]) or known(edits1(word)) or known_edits2(word) or [word]
return max(candidates, key=NWORDS.get)
```

திருத்து என்ற செயல்பாடு (function correct) மூலச் சொல்லுக்கு குறைந்த திருத்து தூரம் இருக்கிற கருதத்தக்கச் சொற்களின் (candidate words) குழுமத்தை தெரிந்தெடுக்கின்றது; அந்த குழுமத்தில் சில தெரிந்த சொற்கள் (known words) இருப்பதுவரை இது நடைபெறும். இது கருத்தில் கொள்ளத்தக்க குழுமத்தை அடையாளம் அடையாளம் கண்டுகொண்டால் அது NWORDS மாதிரியால் மதிப்பிடப்பட்ட மிகக்கூடுதல் P(c) மதிப்புள்ள தனிமத்தைத் தெர்ந்தெடுக்கும்.

மேற்கூறியவாறு தமிழ் எழுத்துப்பிழை மற்றும் இலக்கணப் பிழை திருத்தி உருவாக்கப்பட்டு தட்டச்சு செய்யபட்ட பனுவல்களை திருத்தப் பயன்படுத்தப்படும். இத்தகைய முயற்சி இங்கு ஒரு மாதிரிக்காக மேற்கொள்ளப்பட்டுள்ளது. முன்னர் கூறியது போன்று இவ்வாய்வு ஒரு பரிசோதனை என்ற நிலையிலேயெ அமைகின்றது.

7.6. சுருக்கம்

தமிழில் வெளிவந்த எழுத்துப் பிழை மற்றும் இலக்கணப் பிழை திருத்திகளின் பல்வேறுபட்ட பரிமாணங்கள் குறித்து இங்கு விளக்கப்பட்டது. சொல்லாய்விகளும் சொல்லாளர்களும் பிழை திருத்திகளை உள்ளடக்கி வெளிவரத் தொடங்கின. தொழில் நுட்பவியலார்களும் மொழியியலாளர்களும் இதற்கென்று முக்கியத்துவம் கொடுத்து பல நிலைகளில் எழுத்து மற்றும் இலக்கணத் திருத்திகளை உருவாக்க முயன்றனர். மொழியியல் விதிகளை உள்ளடக்கி திருத்திகளை உருவாக்கச் சிலர் முயன்றனர். இதில் குறிப்பிடத் தகுந்தவர் தெய்வசுந்தரம் அவர்கள். இந்த விதிகளை அறியாத அல்லது விதிகளைப் பற்றி கவனம் செலுத்தாது உடனடியான எழுத்துப் பிழை திருத்தியும் இலக்கணப் பிழை திருத்தியும் உருவாக்க முயன்றவர்களில் பலர் தரவுத்தொகுதியைப் பயன்படுத்தி இத்தகைய பிழை திருத்திகளை உருவாக்க முயன்றனர். எல்லா முயற்சிகளும் முன்னேற்றத்திற்கும் இத்தொழில் நுட்பத்தைப் பற்றி நன்கு புரிந்துகொள்ளவும் வழிவகுத்தன.

தமிழில் எழுத்துப் பிழை திருத்தியும் இலக்கணப் பிழைத் திருத்தியும் உருவாக்குவதற்கான வழிமுறைகளை ஆய்வதற்காக செய்யப்பட்ட ஆய்வு இது ஆகும். இத்தகைய திருத்திகளின் தேவை எப்போதோ உணரப்பட்டு இதற்காண முயற்சிகள் பல மேற்கொள்ளப்பட்டன. இத்தகைய

முயற்சிகள் பல சரியான திசையில் செல்லாததன் காரணமாகக் கைவிடப்பட்டன அல்லது சரியான விளைவுகளை ஏற்படுத்தவில்லை. மொழியியல் அடிப்படையில் இத்தகைய ஆய்வுகள் அமையாதது இதற்கு முக்கியகாரணமாகும். எந்த ஆய்வும் அடிப்படையான தகவலகளை நன்கு தெரிந்த பின்னர்தான் மேற்கொள்ளப்படவேண்டும். எழுத்துப் பிழை மற்றும் இலக்கணப் பிழை திருத்தி உருவாக்க ஆய்வும் இதற்கு விதிவிலக்கல்ல. எவ்வாறு நோயைப் நன்கு புரிந்துகொண்டு சிகிச்சை ஆரம்பிக்கப் படுகின்றதோ அதுபோல் தான் மொழியியல் நுட்பங்களை நன்கு புரிந்தபின் தான் இத்தகைய ஆய்வுகள் மேற்கொள்ளப்படவேண்டும். கணினியின் திறனை மட்டும் நம்பிக்கொண்டு செய்யப்படும் இத்தகைய ஆய்வுகள் குறை உடையதாகவும் பயனின்றியும் தான் அமையும். மொழியியல் மொழியின் இலக்கணத்தை நன்கு ஆய்ந்து தேவையான தகவலகளைத் தருகின்றது. அத்தகைய தகவல்கள் பிழை திருத்த ஆய்வுக்கு முக்கியமானதாகும். இத்தகைய ஆய்வு அடிப்படைகள் தான் திறமைமிக்க திருத்திகளை உருவாக்கத் துணை புரிய இயலும்.

தமிழில் பிழை திருத்திகள் பல உருவாக்கப்பட்டுள்ளன என்பதில் ஐயம் இல்லை. இவைகள் யாவும் எழுத்துப் பிழை திருத்திகள் என்ற நிலையில் தான் அமைகின்றன. இலக்கணப் பிழை திருத்தி என்று கூறத்தக்க கருவிகள் இல்லை என்று கூறலாம். எழுத்துப் பிழை திருத்திகளே இலக்கணப் பிழை திருத்திகளாகவும் உரிமை கோரப்படுகின்றன. இலக்கணத் திருத்தி உருவாக்கம் மிகப்பெரிய செலாகும். கணிப்பொறிக்கு தமிழ் இலக்கணத்தின் நெளிவு சுழிவுகளைக் கற்றுத் தந்தால்தான் இலக்கணத் திருத்தியாகக் கணினி செயலாற்ற் இயலும். தரவுத்தொகுதிகளைக் கொண்டு கணிப்பொறி கற்றல் அணுகுமுறை அடிப்படையில் (machine learning approach) இத்தகைய செயல்பாட்டை எளிதாக இயலும். இதை தரவுத்தொகுதி இயக்க அணுகுமுறை (corpus driven approach) என்றும் புள்ளியியல் அணுகு முறை (statistical approach) என்றும் கூறலாம்.

இரண்டுவிதமான அணுகுமுறைகள் உள்ளதாக முன்னர் பார்த்தோம்: (1) தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை (முற்றுநிலைத் தானியங்கியைப் பயன்படுத்தி) மற்றும் (2) மொழியியல் அடிப்படையிலான அணுகுமுறை (உருபனியல் ஆய்வியைப் பயன்படுத்தி). முதல் அணுகுமுறையில் தவறான சொல் தரவுத்தொகுதியின் மூலம் சரிபார்க்கப்பட்டு சரியான சொல் பரிந்துரைக்கப்படும். இரண்டாவது அணுகு முறையில் தவரான உருபனியல் பகுப்பாய்வு சொல்லின் தவறை வெளிப்படுத்தி சரியான சொல்லைப் பரிந்துரை செய்யும். பரிந்துரையில் முக்கியமாகத் திருத்து தூரம் கணக்கில் கொள்ளப்படும். முன்னர் கூறியது போன்று பிழைகளை விடுபடு பிழை, செருகல் பிழை, பதிலீட்டுப் பிழை, இடமாற்றுப் பிழை என்று நான்காகப் பகுத்தாயலாம் என்று கண்டோம். இந்நான்கில் ஒரு பிழை மட்டும் கடுத்தில் கொள்ளப்பட்ட

சொல்லில் இருந்தால் இதற்கான பரிந்துரை எளிதாக அமையும். அனால் ஒன்றிற்கும் மேற்பட்ட வகை பிழைகள் இருந்தால் பரிந்துரைகள் சிக்கலாக அமையும்.

பிழை ஆய்வை பின்பலமாக வைத்துக்கொண்டு பிழை திருத்தி உருவாக்குவது மிகுந்த பலனை அளிக்கும். இலக்கண விதிகளை வைத்துக்கொண்டு இவ்விதிகள் எவ்வாறு மீறப்படுகின்றன என்ற அடிப்படையில் இலக்கணத்திருத்திகள் உருவாக்குவது சரியான அணுகு முறையாகும். எழுத்துப் பிழை திருத்திகளைப் பொறுத்தவரியில் தமிழ் மொழியின் ஒலியனியல் அமைப்பையும் உருபொலியனியல் விதிகளையும் நன்றாக விளங்கிக்கொண்டு அதன் அடிப்படையில் எழுத்துப் பிழை திருத்தியை உருவாக்குவது நன்மை பயக்கும். இதில் சிக்கலான தகவல் என்னவென்றால் எல்லா விதிகளுக்கும் விதி விலக்குகள் காணப்படும். மொழி பிற மொழிகளிலிருந்து கடன் வாங்குவதன் காரணமாக அதன் தனித்தன்மையான ஒலியனியல் அமைப்பு சிதைவுற வாய்ப்புண்டு. தமிழில் சில எழத்துக்கள் மொழி முதலில் வரும், சில எழுத்துக்கள் மொழி இடையில் வரும், சில எழுத்துக்கள் மொழி இறுதியில் வரும் என்பன போன்ற ஒலியனையல் கட்டுப்பாடுகள் உண்டு. மேலும் மெய்மயக்கங்களின் வருகை முறையும் விதிகளுக்குக் கட்டுண்டு வரும். இத்தகைய மொழி தொடர்பான தகவல்கள் திருத்திகள் உருவாக்கத்திற்குக் கைகொடுக்கும்.

Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

இயல் 8

சொல் மூலங்களை உருவாக்குதல்

8.1அறிமுகம்

சொல்வளங்கள் சொல்லடைவு, சொல்லாட்சி அகராதி, பொருளடைவு, அகராதி, சொற்களஞ்சியம், சொல்லவலை என்பனவற்றை உள்ளடக்கும். இச் சொல்வளங்களை எவ்வாறு கணினியின் உதவியால் உருவாக்க இயலும் என்று இங்கு விரிவாகக் கூறப்படுகின்றது.

8.2.சொல்லடைவு உருவாக்குதல்

சொல்லடைவு தயாரிப்பதையும் கணிப்பொறி மூலம் விரைவில் செய்துவிட இயலும். சொற்களையும் அவற்றோடு தொடர்பான செய்திகளையும் தந்தால் குறைந்த கால நேரத்தில் கணிப்பொறியின் உதவியால் சொல்லடைவு தயாரித்து விடலாம்.

எடுத்துக்காட்டாகச் சங்க இலக்கியங்களுக்குச் சொல்லடைவு தயாரிப்பதை எடுத்துக் கொள்வோம். பின்வரும் அடிகள் சங்க இலக்கியமான புறநானூற்றிலிருந்து எடுக்கப்பட்டவை.

1. இவனியாரென்குவை யாயி னிவனே (புறம் 83.1)
2. தூங்குகையா னோங்க் நடைய (புறம் 22.1)

இவ்வடிகள் சொற்களாகப் பின்வருமாறு பிரித்துத் தரப்படவேண்டும். வழியமைப்புகள் மூலம் இச்சொற்களை அகரவரிசைப்படுத்தி அவை ஒவ்வொன்றும் வரும் இடங்களைக் குறிப்பிட்டுத்தர இயலும்.

ஆயின் (புறம் 83.1)

இயார் (புறம் 83.1)

இவன் (புறம் 83.1)

என்குவை (புறம் 83.1)

ஏ (புறம் 83.1)

ஓங்கு (புறம் 22.1)

கையால் (புறம் 22.1)

தூங்கு (புறம் 22.1)

நடைய (புறம் 22.1)

ஒரு சொல் சங்க இலக்கியத்தில் எத்தனை முறை கையாளப் பெறுகின்றது அல்லது குறிப்பிட்ட ஒரு நூலில் எத்தனை முறை கையாளப் பெறுகின்றது என்பன போன்ற தகவல்களை வழியமைப்புகள் மூலம் பெறலாம். தமிழ்ப்பல்கலைக்கழக அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்திச் சங்க இலக்கியங்களுக்கான ஒரு சொல்லடைவைத் தயாரித்துள்ளன.

8.3.சொல்லாட்சி அகராதி மற்றும் சொற்பொருளடைவு உருவாக்குதல்

சொல்லாட்சி அகராதி (Concordance) ஒரு பனுவலில் இடம் பெறும் சொற்களை அகரவரிசைப்படுத்தித் தருவதோடு அச்சொற்கள் நூலில் இடம்பெறும் இடங்கள், எத்தனை முறை இடம்பெறுகின்றன என்ற கணக்கீடு, எந்தெந்தச் சூழலில் பயன்படுத்தப்படுகின்றன என்பன போன்ற செய்திகளைத் தரும். சொற்பொருளடைவு (Compendium) சொல்லாட்சி அகராதிகளினின்றும் ஒருபடி மேற்சென்று சொற்கள் வரும் இடங்களையும் சூழல்களையும் பொருள் அடிப்படையில் பிரித்துத் தரும். கணிப்பொறியைப் பொறுத்தவரையில் இரண்டையும் ஒன்றில் அடக்கித் தந்துவிடலாம்.

புறநானூற்றிலிருந்து எடுக்கப்பட்ட ஒரு சொல்லின் சொற்பொருளடைவின் மாதிரி கீழே தரப்பட்டுள்ளது.

அணி

1. அழகு

1.1 அழகு தாழ்ந்த அணி (புறம் 57.1)

1.2 அணிகொள் தேர் (புறம் 68.14)

. . .

. . .

. . .

2. அணிகலன்

2.1. இழையணிபொலிந்த . . . . . .

(புற, 81.1)

. . .

. . .

. . .

3. ஒப்பனை

3.1. ஆடு கோட்பாட்டுச் சேரலாதன் போந்தைப் பொழியின் கண் அணியல் (படி 51.9)

4. அணிமை

5. வரிசை

6. திரட்சி

7. ஆடை

8. ஒப்புமை

9. சூழ்தல்

10. பக்கரை (சேணம்)

சொல்லடைவு தயாரிக்க முன்னர்க் கூறியது போல் செய்யுளின் அடிகளைச் சொற்களாகப் பிரித்துத் தரவேண்டும். இத்துடன் சூழல் அடையாளப் படுத்தப்பட வேண்டும். வழியமைப்புகள் மூலம் நாம் சொல்லாட்சி அகராதியையோ சொற்பொருடைவையோ தயாரிக்க இயலும். தமிழ்ப்பல்கலைக்கழக் அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்திச் சங்ககால இலக்கியங்களுக்கு இத்தகைய சொல்லாட்சி அகராதியும் சொற்பொருளடைவும் தயாரித்துள்ளன.

8.4. அகராதி உருவாக்குதல்

கணினியைப் பயன்படுத்தி, பலர் உழைப்பால் பல்லாண்டுகாலம் செய்யப்படும் அகராதித் தயாரித்தலை மிகக் குறுகிய காலத்தில் குறைந்த ஆட்களின் சக்தியால் செய்து முடித்துவிடலாம். கணிப்பொறியில் பயன்படுத்தத்தக்க விதத்தில் மின் அகராதிகளையும் தயாரிக்கலாம். இத்தகைய அகராதி கணிப்பொறி வழி மொழி கற்பித்தல், மொழிபெயர்த்தல் போன்ற செயற்பாடுகளுக்குப் பெரிதும் பயன்படும்.

8.4.1.தமிழில் கணிப்பொறி வழி அகராதி உருவாக்குதல்

மொழி கற்றலில் கணிப்பொறியின் பங்கு மகத்தானது. ஒரு மொழியின் சொல் அதன் இலக்கணக் கூறு, பொருள், பயன்பாடு மற்றும் வாக்கியத்தில் அதன் உபயோகம் போன்றவைகளை அகராதி விளக்குகிறது. மேலும் ஒரு சொல்லுக்கு நிகரான பிறமொழி பதங்களையும் இருமொழி அகராதிகள் விளங்குகின்றன. அகராதி தயாரிக்கும் பணி எளிதானதல்ல. மிகுந்த நேரமும், அதிக உடல் உழைப்பும் பொருட்செல்வம் ஆகும். ஒரு அகராதியை உருவாக்க நீண்டகாலம் தேவைப்படும். எனவே அகராதி உருவாக்க கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன.

8.4.2. அகராதி உருவாக்கும் படிநிலைகள்

பொதுவாக அகராதி ஒன்றை தயாரிப்பதற்கு தேவைப்படும் படிநிலைகள்:

1. செய்திகளை சேகரித்தல், ஆராய்ந்து வகைப்படுத்துதல் அகர வரிசைப் படுத்துதல் என்பது முதற்கட்ட நிலைகளாகும்.

2. சொல்லுக்கான இலக்கணக்கூறு ஆராய்தல், பொருள் எழுதுதல், பிறமொழி சொற்களை எழுதுதல், சொல்லின் உபயோகத்தை விளக்குதல் இவைகளை இரண்டாம் நிலைப் பணியாகும்.

இதில் முதல் கட்ட பணி எளிமையாக கணிப்பொறி படுத்த இயலும். இரண்டாம் கட்டப்பணியை மனித முயற்ச்சியாலேயே செய்ய இயலும்.

8.4.3. கணிப்பொறி வழி அகராதி: வரலாற்றுச் சுருக்கம்

மாத்தூஸ் அர்னால் (Mathews Arnald) என்ற ஆங்கில கவிஞர் தன் கவிதைகளுக்கான சொல்லடைவை முதன் முதலாக கணிப்பொறி வழி உருவாக்கம் செய்தார். இதற்கு IBM-704 என்ற கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறி வழி சொல்லடைவினை முதன் முதலாக பேராசிரியர் ஸ்டிபன் (Stephen) உருவாக்கினார். இதனால் தொடர்ந்து Oxford அகராதி உருவாக்க சொல்லடைவு (Concordance) கணிப்பொறி வழி உருவாக்கப்பட்டது. இதைத் தொடர்ந்து ஆங்கிலம் மற்றும் பிற மொழிகளில் அகராதி உருவாக்க கணிப்பொறிகள் ஆங்காங்கே பயன்படுத்தப்பட்டன. தமிழ்ப்பல்கலைக்கழகத்தில் சங்க இலக்கியத்திற்கான' அகராதி உருவாக்க கணிப்பொறி பயன்படுத்தப்பட்டது. இதுவே முதல் முயற்சி ஆகும். கிரியா என்னும் தற்கால அகராதி கணிப்பொறி வழி உருவாக்கப்பட்டுள்ளது. இந்த அகராதி அச்சு வடிவிலும், குறுந்தகடுகளிலும் கிடைக்கப் பெறுகின்றன.

8.4.4.கணிப்பொறி வழி அகராதி: முறைமைகள்

கணிப்பொறி வழி அகராதி உருவாக்குவதற்கு கீழ்கண்ட முறைமைகள் பின்பற்றப்படுகின்றன.

1. செய்திகளை உள்ளீடு செய்தல்
2. செய்திகளை வார்த்தைகளாகப் பிரித்தல்

இந்நிலையில் ஒரு வாக்கியத்திலிருந்து சொற்கள் தனியாக பிரிக்கப்பட்டு பின்வரும் பட்டியல் உருவாக்கப்படுகிறது.

சொல்	வரி எண்
WORD	LINE NUMBER

1. மேற்கண்ட பட்டியலில் சொற்கள் அவைகள் இடம் பெறும் வரிசையில் இடம் பெறுகின்றன. மேலும் ஒரு சொல் உரையில் பல இடங்களில் பயனபடுத்தப்பட்டிருந்தால் அத்தனை முறை அச்சொல் இப்பட்டியலில் இடம்பெறும்.

2. மேற்கண்ட பட்டியல் அகரவரிசைப் படுத்தப்படுகிறது. இது மேற்கண்ட பட்டியலில் சொற்கள் இடம்பெறும் வரிசை முறையை மாற்றுகிறது. ஒரு சொல் பல முறை இடம் பெற்றிருப்பினும் அது அடுத்தடுத்து வரிசைமுறை பெறுகிறது.

3. இவ்வாறு அகரவரிசைப் படுத்தப் பட்டுள்ள பட்டுள்ள பட்டியலிலிருந்து கீழ்கண்ட பட்டியல் உருவாக்கப்படுகிறது.

சொல் (WORD)	நிகழ்வெண்ணிக்கை FREQUENCY	சொல் இடம்பெறும் இடம் LIST OF LINE NUMBER

இப்பட்டியலின் உதவிக் கொண்டு பின்வரும் பட்டியல் உருவாக்கப்படுகின்றது.

சொல் WORD	நிகழ்வெண்ணிக்கை FREQUENCY	சொல் இடம்பெறும் வரிசை எண் LIST OF LINE NUMBER	இடம் பெறும் வரி CITATION

மேற்கண்ட பட்டியலிலிருந்து ஒரு சொல் உரையில் எத்தனை முறை பயன்படுத்தப்பட்டுள்ளது என்பதை அறியலாம். அச்சொல் இடம்பெறும் வரி தெரிவதால் அதனைக்கொண்டு அச்சொல்லின் இலக்கணக்கூறு பொருள் உபயோகம் போன்றவைகளை ஆராய்ந்து எழுதலாம். இதன்மூலம் விவரங்கள் துல்லியமாகவும் விரைவாகவும் கிடைக்கப்பெறுகின்றன.

8.4.5. அகராதி உருவாக்குவதில் இயற்கை மொழியாய்வு

கணிப்பொறி வழி அகராதி தயாரிக்கப்படும்போது கிடைக்கப்படும் இறுதி படிவம் சொல்லடைவு படிவம் போன்று இருக்கும். இதில் பெயர் சொற்கள் மற்றும் வேற்றுமை உருபுகளுடன் கூடிய பெயர் சொற்கள் மற்றும் வேற்றுமை உருபுகளுடன் கூடிய பெயர் சொற்கள் தனித்தனி சொற்களாகவே இடம்பெறும். உதாரணமாக கணிப்பொறி, கணிப்பொறியை, கணிப்பொறியில், கணிப்பொறியால், கணிப்பொறிக்கு போன்ற சொற்கள் தனிச்சொற்களாகவே இடம்பெறும். இவைகளை சொல் வடிவங்கள் என மொழியியலாளர்கள் கூறுகின்றனர். ஆனால் அகராதியில் இவைகள் தனிச்சொற்களாக இடம்பெறும் (அல்லது) இடம்பெறாது. கணிப்பொறி என்ற பெயர்ச்சொல் மட்டுமே அகராதியில் இடம்பெறும். இது போன்ற அகராதி சொற்களை மொழியியலாளர்கள் lexmes என்கின்றனர். இது போன்று படி என்ற வினைச் சொல்லுக்கு படித்தான், படித்தார், படித்தது போன்ற எண்ணற்ற சொல் வடிவங்கள் இருப்பினும் படி என்பதே அகராதிச் சொல்லாகும். கணிப்பொறி வழி சொல்லடைவு தயாரிக்கும்போது சொல் வடிவங்களே இடம்பெறும். இதிலிருந்து அகராதிச் சொற்களை மனித முயற்சியினாலேயே மேற்கொள்ள வேண்டும். இப்பணியையும் முழுமையாக கணிப்பொறிப்படுத்த உருபனியல் ஆய்வு மேற்;கொள்ள வேண்டும். இவ்வாறாக சொல் வடிவத்திலிருந்து (word forms) உருபனியல் பகுப்பாய்வு மூலம் அகராதி சொற்களை (lexems/dictionary words) உருவாக்கும் முறையை மொழியியலாளர்கள் lemmantisation என்று அழைக்கின்றனர். இம்முறையை கணிப்பொறிப்படுத்த முயற்சிகள் மேற்கொள்ளப்பட்டு வருகிறது. இதற்கு இயற்கை மொழி ஆய்வு உதவுகிறது.

8.4.6. முடிவுரை

கணிப்பொறி வழி அகராதி தயாரிக்கும் படிநிலைகளில் கணிப்பொறி படுத்த இயலும் நிலைகளையும், முறைகளையும் சுருக்கமாகப் பார்த்தோம். இம்முறைகளில் அகராதியில் மொழி ஒரு தடையில்லை. தமிழ் அகராதியோ, ஆங்கில அகராதியோ, பிறமொழி அகராதிகளையோ மேற்கண்ட முறைகளைப் பின்பற்றி உருவாக்கலாம். அந்தந்த மொழிகளில் அகர வரிசைப்படுத்தப் படும் மென்பொருள்களை வடிவமைக்க வேண்டும்.

அகராதி தயாரிப்பதற்கான சில மென்பொருள்கள் வெளிவந்துள்ளன. (எடுத்துக்காட்டாக Tool Box என்ற மென்பொருள் அகராதி தயாரிக்கப் பெரிதும் உதவும்.) அவற்றைப் பயன்படுத்தி அகராதிகள் தயாரிக்கலாம். DATA BASE, DATE BASE MANAGEMENT SYSTEMS போன்றவை அகராதி தயாரித்தலுக்குப் பெரிதும் உதவும். பின்வரும் எடுத்துக்காட்டு ஃபாக்ஸ்புரோ போன்ற வழிமுறைகளைப் பயன்படுத்தி அகராதி தயாரித்தலின் எளிய விளக்கமாகும்.

அகராதியின் ஒவ்வொரு சொற்பதிவிலும் தலைப்புச் சொல், எழுத்துப்பெயர்ப்பு, சொல் வகைப்பாடு, பல்பொருண்மைகள் போன்ற தகவல்கள் அடங்க வேண்டும் என்பதன் அடிப்படையில் தரவு மையத்தை உருவாக்க வேண்டும். கீழ்வரும் எடுத்துக்காட்டில் முக்காற்புள்ளிக்கு இடப் பக்கம் வருவனவற்றைப் பீல்டுகளின் பெயர்களாகவும் முக்காற்புள்ளிக்கு வலப்பக்கம் வருவனவற்றை அவற்றில் நிரப்பப்பட வேண்டிய செய்திகளாகவும் எடுத்துக்கொள்ளலாம். ஒவ்வொரு சொற்பதிவும் ஒரு பதிவாக அமையும்.

தலைப்புச் சொல்:ஏவு

எழுத்துப்பெயர்ப்பு: eevu

சொல் வகைப்பாடு: செ.பொ.குன்றா.வி. (செயப்படுபொருள் குன்றா வினை)

பொருள் 1: தூண்டுதல்

எடுத்துக்காட்டு 1: குண்டர்களை ஏவி எங்களைப் பயமுறுத்தினான்.

பொருள் 2 : 2. கட்டவிழ்த்துவிடல்

எடுத்துக்காட்டு 2 : அடக்கு முறையை ஏவிவிட்டு மக்களின் போராட்டத்தை அடக்கி விட இயலாது.

பொருள் 3 : அனுப்புதல், செலுத்துதல்

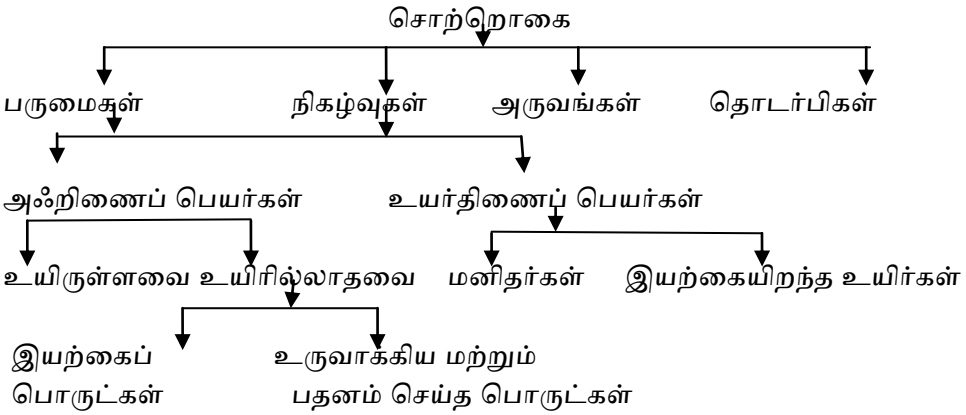
எடுத்துக்காட்டு 3 : ஏவுகணைகளை ஏவிப் பலத்த சேதம் விளைவித்தனர்.

வழியமைப்புகளைப் பயன்படுத்திச் சொற்பதிவைப் பின்வருமாறு அகராதி வடிவில் தரலாம்.

ஏவு. eevu செ. பொ. குன்ற.வி. 1. தூண்டுதல்; குண்டர்களை ஏவி எங்களைப் பயமுறுத்தினான். 2. கட்டவிழ்த்துவிடல்; அடக்கு முறையை ஏவிவிட்டு மக்களின் போராட்டத்தை அடக்கிவிட இயலாது. 3. அனுப்புதல், செலுத்துதல்; ஏவுகணைகளை ஏவிப்பலத்த சேதம் விளைவித்தனர்.

8.5.சொற்களஞ்சியம் உருவாக்குதல்

சொற்களஞ்சியம் (thesaurus) தயாரிப்பதற்குக் கணிப்பொறியைப் பயன்படுத்தலாம். ஒத்தல் (similarity), உள்ளடங்கல் (Inclusion), மேலுறல் (Overlapping), வேறுபடல் (Disjunction) என்ற தர்க்கப் பொருண்மை உறவுகளின் அடிப்படையில் சொற்களை அவற்றின் படிநிலை அமைப்பும் பொருண்மையும் வெளிப்படும் விதத்தில் பொருண்மைக் களங்களாகப் பகுத்துத் தருவது சொற்களஞ்சியமாகும். கணிப்பொறி உதவியுடன் இச்சொற்களஞ்சியத்தைக் குறைந்த உழைப்பில் குறுகிய காலக் கட்டத்தில் உருவாக்கிவிடலாம். கணிப்பொறியில் பயன்படுத்தத்தக்க விதத்தில் மின் சொற்களஞ்சியத்தை உருவாக்கலாம். இத்தகைய மின்சொற்களஞ்சியம் மொழிபெயர்த்தல், மொழி கற்பித்தல், சொற்பொருண்மை மயக்க நீக்கம், தகவல் மீட்பு போன்ற செயற்பாடுகளுக்குப் பெரிதும் பயன்படும். தமிழுக்கான மின் சொற்களஞ்சியம் ஒன்றைத் தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து உருவாக்கியுள்ளன (இராசேந்திரன் மற்றும் பாஸ்கரன் (2006). இதன் தரவுத்தளம் இராசேந்திரனின் (2001) தற்காலத் தமிழ்சொற்களஞ்சியத்தைப் பயன்படுத்தி உருவாக்கப்பட்டுள்ளது. இச்சொற்களஞ்சியத்தின் தரவுதளம் நைடாவின் (1975) பொருண்மைக்கூறாய்வு அடிப்படையில் உருவாக்கப்பட்டது. அவர் பைபிள் மொழிபெயப்புக்காக சொற்களை மெய்படுத்தம் செய்யும் கருத்துருக்களைப் படிநிலை அமைப்பாக பொருட்புல வகைப்பாடு செய்து தந்துள்ளார். (இதில் தரப்பட்டுள்ள சொற்றொகைப் பொருண்மை அமைப்பு காட்சி மூலப்பொருண்மையியல் உருவாக்கத்தின் கீழ் விளக்கமாகத் தரப்பட்டுள்ளது.) கீழ்வருவது இதன் சுருக்கமான கிளைப்படமாகும்.



நைடாவைப் (1975) பின்பற்றி சொற்களால் மெய்படுத்தம் செய்யப்படும் கருத்துருக்கள் (concepts) பருமைகள் (entities), நிகழ்வுகள் (events), அருவங்கள் (abstracts), தொடர்பிகள் (relationals) என நன்காகப் பகுக்கப்பட்டுள்ளன. பருமைகள் அருவமல்லாப் பருமைப்பெயர்களை உள்ளடக்கும். நிகழ்வுகள் வினைகளையும் வினைப்பெயர்களையும் உள்ளடக்கும்; அருவங்கள் அருவப்பெயர்களையும் பெயரடைகளையும் வினையடைகளையும் மெய்ப்படுத்தம் செய்யும்; தொடர்பிகள் பின்னுருபுகளையும் பிற செயற்பாட்டுச் சொற்களையும் (functional words) மெய்ப்படுத்தம் செய்யும். இப்பொருண்மைக் களங்களில் உட்படுத்தப்படும் சொற்கள் ஒருபொருபொருள் பன்மொழியம்(synonymy), உள்ளடங்கு மொழியம்-உள்ளடக்கு மொழியம் (hyponymy-hypernymy), சினைமொழியம்-முழுமொழியம் (meronymy-holonymy), எதிர்மொழியம் (antonymy) போன்ற பொருண்மை அல்லது சொல்சார் உறவுகளால் தொடர்பு படுத்தப்பட்டுள்ளன. பயனர் இடைமுகத்தின் மூலம் வேண்டிய சொல்லை அதில் உள்ளீடு செய்து அதன் ஒருபொருள் பன்மொழிகளையோ உள்ளடங்கு மொழியையோ உள்ளடக்கு மொழியையோ சினை மொழியையோ முழுமொழியையோ எதிர்மொழியையோ தரவுத்தளத்திலிருது பெறலாம்.

8.6.சொல்வலை உருவாக்குதல்

இது ஒரு இணையதளத் தரவு மையமாகும் (Online Data Base). இது சொற்களஞ்சியம் மற்றும் அகராதி என்ற இரண்டின் பண்புகளையும் கொண்டு மிகுந்த திறன் உடையதாய் அமைந்துள்ளது. சொற்கள் தமக்கிடையே உள்ள பொருண்மை உறவுகளால் வலை போல் பின்னப்பட்டிருப்பதை வெளிப்படுத்துவதாய் இது அமையும். சொல்வலையின் மிகப் பரந்த நோக்கம் சொல் தகவல்களைச் சொல்வடிவுகள் அடிப்படையில் அல்லாமல் சொற்பொருண்மை அடிப்படையில் ஒழுங்குபடுத்துவதாகும். இந்த அடிப்படையில் சொல்வலை அகராதியிலிருந்து

வேறுபட்டுச் சொற்களஞ்சியத்தையும் உள்ளடக்கும். ரொஜஸ்ட் தெசாரஸ் போன்ற சொற்களஞ்சியங்களும் தமிழுக்கு இராசேந்திரனால் (2001) உருவாக்கப்பட்ட தமிழ்ச் சொற்களஞ்சியமும் சொல்வலை பின்னுவதற்கு முன்னோடியாய் அமையும். இவையெல்லாம் புத்தக வடிவில் அமைந்தவை. அகரவரிசையில் அமைந்த சொற்களஞ்சியங்களில் சிக்கல் என்னவென்றால் அவற்றில் தேவைக்கு அதிகமான பதிவுகள் காணப்பெறும். இரு சொற்கள் ஒருபொருள் பன்மொழிகள் என்றால் கூட அகரவரிசைக்காக இரண்டுமுறை நோக்கீடு செய்துதான் தேவையான தகவல்களைப் பெரும்பாலும் பெற இயலும். முதல்முறை அகரவரிசையில் அமைந்த சொல்லடைவு மூலம் குறிப்பிட்ட சொல் எப்பொருண்மைத் தலைப்பில் வருகிறது என்பதைத் தெரிந்துகொண்டு இரண்டாவது முறை நோக்கீடு செய்துதான் வேண்டிய தகவலைப் பெற இயலுகின்றது. இதனால் தேடும் நேரம் இரண்டு மடங்காக உயரும். ஆனால் கணிப்பொறியின் வழி தேடலை விரைவாகச் செய்ய இயலும். சொல் வலை மின்சொற்களஞ்சியத்தைவிட ஒரு படி மேலானது. தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையில் இராசேந்திரனை முதன்மை ஆய்வாளராகக் கொண்டு இந்திய அரசின் தகவல்தொடர்பு மற்றும் மின்னணு அமைச்சின் நிதி நல்கையின் கீழ் 2006-2009 என்ற காலகட்டடத்தில் திராவிடச் சொல்வலை உருவாகத்தின் ஒருபகுதியாகத் தமிழ்ச் சொல்வலை உருவாக்கம் தொடங்கப்பட்டது முதல் கட்டத்தில் 15000 ஒருபொருள் பன்மொழியக் குழுங்களும் இரண்டாம் கட்டத்தில் 15000 ஒருபொருள் பன்மொழியக் குழுமங்களும் உருவாக்கப்பட்டு TDIL இணைய தளத்தில் பதிவேற்றம் செய்யப்பட்டுள்ளன.

சுருக்கமாக, சொற்களை அவற்றிற்கு இடையே காணப்படும் உறவுகள் அடிப்படையில் பிணைக்கும் ஒரு ஒழுங்குமுறை சொல்வலை (wordNet) என வரையறை விளக்கம் செய்யலாம். சொல்வலையின் அடிப்படை உருவாக்கக் கூறு ஒருபொருள்பன்மொழியக் குழுமங்களாகும் (synsets). ஒருபொருள்பன்மொழியக் குழுமங்கள்தாம் சொல் உறவுகளாலும் பொருண்மை உறவுகளாலும் தொடர்புபடுத்தப்பட்டுள்ளன. சொல்வலையின் பின்னணியில் மூலப்பொருண்மை ஆய்வு (ontology) அமையும்.

மொழி ஒரு ஒருங்கிணைந்த, ஒன்றிணைக்கப்பட்ட ஒழுங்குமுறையாகச் செயல்படுகின்றது. அடிப்படையில் மொழி உலகை மனித மூளையுடன் தொடர்புபடுத்துகின்றது. புலனால் உலகைக் கருத்துருவாக்கம் செய்ததன் விளைவாக மொழியின் அடிப்படை அலகுகளான சொற்கள் உருவாக்கப்பட்டன. அவை மூளைக்கு உள்ளீடுகளாக அமைந்து கருத்துப் பரிமாற்றத்தின் வெளியீடாக அமைந்தது. எனவே சொல்வலையை ஒரு மன/உள அகராதி (Mental dictionary) எனக் கூறலாம்.

சொல்வலையின் முக்கியப் பயன்பாடு சொற்கள் வழியாகத் தகவல்களைப் பெற உதவும் ஒழுங்குமுறையாக (Information Access system) அமைவதாகும். சொல்வலை பல பொருண்மை உறவுகளில் ஒரு கருத்துருவை மற்றொரு கருத்துருவுடன் இணைப்பதன் காரணமாக அது இயல்பாகவே ஒரு எடுத்துக்காட்டான சொல் பெறும் ஒழுங்குமுறையாக (Word access system) மாறியுள்ளது. சொல்வலை பல நடைமுறைப் பயன்பாடுகள் கொண்ட ஒரு சொல் தரவு மையமாகும் (lexical database).

சொல்வலையின் பயன்பாடுகளில் ஒன்று மொழிபெயர்ப்பாகும். சொல்வலை ஒரு கருத்துருவை (concept) மற்றொரு கருத்துருவுடன் சொல் உறவுகளாலும் (lexical relations) பொருண்மை உறவுகளாலும் (semantic relations) தொடர்புபடுத்துவதோடு மட்டுமல்லாமல், ஒரு சொல்லின் சூழல் அடிப்படையிலான பொருண்மை வேறுபாடுகளையும் தருகின்றது. சொல்வலை பொருண்மையைச் சூழலுடன் தொடர்புபடுத்தும் தன்மை கொண்டிருப்பதன் காரணமாக அது ஒரு சொல்லின் பொருண்மைகளைச் சூழல் அடிப்படையில் வேறுபடுத்த இயலும்.

8.6.1. இந்தோசொல்வலை

சொல்வலைகள் உலகமுழுவதும் பல மொழிகளுக்காக உருவாக்கப்பட்டுள்ளன. எடுத்துக்காட்டாக ஆங்கிலம், செருமன், பிரஞ்சு, இத்தாலி போன்ற ஐரோப்பிய மொழிகளுக்குச் சொல்வலைகள் உருவாக்கப்பட்டுப் பயன்பாட்டில் உள்ளன. இவைகளை ஒன்றிணைக்கும் பன்மொழிய யூரோ சொல்வலையும் (Euro WordNet) உருவாக்கப்பட்டுள்ளது. யூரோ சொல்வலையைப் பின்பற்றி இந்திய மொழிகளுக்கும் இந்தோ சொல்வலை (Indowordnet) உருவாக்கும் செயல்பாடு தொடங்கப்பட்டுச் செயல்பட்டு வருகின்றது. இந்திக்கான சொல்வலை தான் இந்தியாவில் உருவாக்கப்பட்ட முதல் சொல்வலையாகும். தற்போது மராத்தி, குஜராத்தி, ஒரியா, பெங்காளி, அசாமீஸ், கொங்கணி, பஞ்சாபி, போடோ, மணிப்புரி, சமஸ்கிருதம் ஆகிய மொழிகளுக்குச் சொல்வலைகள் உருவாக்கும் செயல்பாடு நடைபெற்றுவருகின்றது.

8.6.2. திராவிடச் சொல்வலை

இந்திய அரசின் மின்னணு மற்றும் தகவல் தொழில் நுட்ப நிறுவனத்தின் (Department of Electronics and Information Technology) நிதி நல்கையின் கீழ் திராவிட மொழிகளுக்கான சொல்வலைகள் உருவாக்கும் செயல்பாடும் தொடங்கப்பட்டு செயல்பட்டு வருகின்றது. தமிழ் (தமிழ்ப் பல்கலைக் கழகம்), மலையாளம் (அமிர்தா பல்கலைக் கழகம்ஆ), தெலுங்கு (திராவிடப் பல்கலைக் கழகம்), கன்னடம் (மைசூர் பல்கலைக் கழகம்) ஆகிய திராவிட மொழிகளுக்குச் சொல்வலைகள் உருவாக்கும் செயல்பாடு கடந்த சில ஆண்டுகளாக நடைபெற்றுவருகின்றது.

திராவிட மொழிகள் இலக்கண அடிப்படையிலும் சொற்றொகுதிகள் அடிப்படையிலும் இந்தோ ஆரிய மொழிகளிலிருந்து வேறுபடுவதன் காரணமாகத் திராவிடச் சொல்வலை உருவாக்கம் முக்கியமானதாகக் கருதப்படுகின்றது. தற்போது இந்தி சொல்வலையை அடிப்படையாகக் கொண்டே பிற இந்திய மொழிகளுக்குச் சொல்வலை உருவாக்கும் செயல்பாடு நடைபெற்று வருகின்றது. பின்வருவன திராவிடச் சொல்வலை உருவாக்குவதன் நோக்கங்களாகும்

- திராவிடமொழிகளுக்கு மிக விரிவான, மிகச் சிறந்த பன்மொழிய பொருண்மைசார் சொல் தரவுமையம் உருவாக்குவது.
- மொழி வலையமைப்பைத் தொடர்புபடுத்தி மொழிச் சுதந்திரமான பொருண்மை கருத்துருக்களை உருவாக்குவது.
- எல்லாத் திராவிட மொழிகளுக்கும் பொருண்மைசார் தகவல்களை வகைப்படுத்துவதைத் தரப்படுத்துவது மற்றும் பயன்பாடுகளுக்கு மூலவளங்களை உருவாக்குவது.

8.6.3. யூரோ சொல்வலை

சொல்வலையைப்பற்றிக் கூறுகையில் யூரோ சொல்வலையைப் பற்றிக் கூறாது செல்ல இயலாது. ஏனென்றால் இது தான் முதன் முதலில் நடை முறைப்படுத்தப்பட்டு இணையத்தில் பயன்பாட்டில் உள்ளது. யூரோ சொல்வலையானது சொல் உறவுகளால் தொடர்புபடுத்தப்பட்ட சொற்களின் பன்மொழியச் சொல் வலையைத் (multi lingual wordtnet) தருகின்றது. இது மொழிகளுக்கு இடையிலான தேடலைச் சாத்தியமாக்குகின்றது. சொல்வலையானது பெயர்கள், வினைகள், வினையடைகள், பெயரடைகள் இவை பற்றிய செய்திகளைக் கொண்டிருக்கின்றது. மேலும் இவை ஒருபொருள்பன்மொழியக் குழுமக் (synset) கருத்துருவைச் சுற்றி ஒழுங்கமைக்கப்பட்டுள்ளது. யூரோ வலையின் முக்கிய நோக்கம் என்னவென்றால் சிக்கனமான முறையில் பலமொழிகளுக்கிடையே (English, Dutch, Italian, Spanish, German, French, Estoninan, Czech) மிக விரிவான மற்றும் தரம் வாய்ந்த பன்மொழியத் தரவு மையத்தை (Data base) உருவாக்குவதாகும். யூரோநெட்டினால் உருவாக்கப்பட்ட அறிவுவளம் பல தரப்பட்ட எல்லைக்குட்பட்ட பயன்பாட்டாளாரால் பயன்படுத்தப்படுகின்றது.

8.6.4. தமிழ்ச் சொல்வலை

தமிழ்ச் சொல்வலையை (Rajendran 2009, 2010, 2016) மையமாகக்கொண்டு சொல்வலை உருவாக்குவது பற்றியும் அதன் பரிமாணங்கள் பற்றியும் இங்கு விளக்கப்பட்டுள்ளது.

8.6.4.1 அகராதியும் சொல் வலையும்

அகராதியானது சொல்லின் எழுத்துவடிவு, பொருள், பயன்பாடு இவற்றைக் காணப் பொதுவாய்ப் பயன்படுத்தப்படுகின்றது. ஆனால் சொல் வலை ஒரு பொருண்மை அகராதியாகும் (semantic lexicon). சொல்வலை சொற்களுக்கு வர்ணனையும் விளக்கமும் மாதிரி எடுத்துக்காட்டுகளும் தருமும். சொற்களுக்கு இடையே உள்ள பொருண்மை உறவுகளை மிகத்தெளிவாகப் பயன்படுத்தக்க விதத்தில் கூறும். உச்சரிப்பு, சொல்லாக்க உருபனியல் செய்திகள், சொல்மூலம் அல்லது வரலாறு இவற்றைத் தராது.

8.6.4.2 சொற்களஞ்சியமும் சொல் வலையும்

முதல் சொற்களஞ்சியம்/பொருட்புல அகராதி (Thesaurus) ராஜெஸ்ட் என்பரால் உருவாக்கப்பட்டது. இது பொருண்மை ஒழுங்கமைப்பைக் கொள்கையாகக் கொண்டு உருவாக்கப்பட்டுள்ளது. கருத்துரு அடிப்படையில் அமைக்கப்பட்டது. பயன்பாட்டாளர் தங்கள் மனதிலுள்ள கருத்துருவைக் கொண்டு சொற்களைக் கண்டுகொள்ள இது உதவும். சொல்வலையும் சொற்களஞ்சியம் போல அமையும். இதன் கட்டுமான அலகு ஒருபொருள்பன்மொழியக் குழுமம் ஆகும். ஒருபொருள்பன்மொழியக் குழுமத்தில் ஒரு கருத்துருவை வெளிப்படுத்தும் எல்லாச் சொற்களும் குழுமம் செய்யப்பட்டிருக்கும். சொல்வலையைப் பயன்படுத்துபவர் மனதிலுள்ள கருத்துருவைக் கொண்டு சொற்களைக் கண்டு கொள்ள இயலும். ஒருபொருள்பன்மொழியக் குழுமங்கள் உள்ளடங்கு-உள்ளடக்கு உறவு (honymy-hypernymy), சினை-முதல் உறவு (part-whole or meronymy-holonymy), உட்படுத்து உறவு (entailment) போன்ற உறவுகளால் இணைக்கப்பட்டுள்ளன. சொல்வலை கருத்து நிலையையும் சொல் நிலையையும் தெளிவாகப் பிரிக்கின்றது. இந்த வேறுபாடு சொல் உறவுகளுக்கும் கருத்துக்களுக்கும் உள்ள வேறுபாட்டில் பிரதிபலிக்கின்றது. சொற்களஞ்சியத்திற்கு அப்பாற்பட்டு சொல்வலையில் சொற்களுக்கும் கருத்துருக்களுக்கும் உள்ள உறவுகள் வெளிப்படையாகக் கூறப்பட்டு அடையாளப்படுத்தப்பட்டுள்ளன. பயன்படுத்துபவுபவர் உளவுகளைத் தெரிந்தெடுத்து, ஒரு கருத்துருவிலிருந்து மற்றொரு கருத்துருவுக்குச் சென்று கருத்துரு வெளியிடத்தில் வலம் வர இயலும்.

8.6.4.3 சொல் வலையில் உறவுகள்

சொல்வலை கருத்துரு-பொருண்மை உறவுகளுக்கும் சொற்களை இணைக்கின்ற சொல் உறவுகளுக்கும் வேறுபாடு காட்டுகின்றது. ஆனால் சொல் நோக்கில் கவனம் செலுத்துபவர் முதன்மையாகச் சொல் உறவுகளைப் பயன்படுத்துவர். சொல்வலை பொருண்மை உறவுகளால் ஒழுங்கமைக்கப்பட்டுள்ளது. சொல் உறவுகள் ஒருபொருள்பன்மொழியக் குழுமங்களுக்கு

இடையே உள்ள காட்டிகளாகும். சொல்வலை வேறுபட்ட தொடரியல் வகைப்பாடுகளில் அடங்கும் சொற்களை இணைக்கும் அடுக்கு உறவுகளைக் கொண்டிருக்கவில்லை. நான்கு முக்கியத் தொடரியல் வகைப்பாடுகளான பெயர், வினை, பெயரடை, வினையடை என்பன தனித்தனியாகத் தரப்பட்டுள்ளன. பெயர்கள் படிநிலை அமைப்புகளாகவும் வினைகள் வேறுபட்ட உட்படுத்து உறவுகளாலும் (entailment relation) பெயரடைகளும் வினையடைகளும் N-பரிமாண உயர்வெளியிடத்திலும் ஒழுங்கமைக்கப்பட்டுள்ளன.

8.6.4.4 பெயர்ச்சொல்வலை

பெயர்ச்சொல் வலையில் அடிப்படை பொருண்மை உறவு ஒருபொருள்பன்மொழிய உறவாகும். ஒருபொருள்பன்மொழியக் குழுமங்கள் அடிப்படைக் கட்டுமான அலகுகளாகும். ஒருபொருள்பன்மொழியக் கருத்துச்சாயல் என்பது எல்லாச் சூழல்களிலும் ஒன்றையொன்று இடம்பெயர்த்தல் என்பதை உட்படுத்தாது. இடம்பெயர்த்தல் என்ற அளவியால் இயற்கை மொழிகளுக்குக் குறைந்த எண்ணிக்கையிலான ஒருபொருள்பன்மொழிகள் மட்டுமே வரும். சொல்வலையில் ஒருபொருள்பன்மொழிகளைக் குறைந்தது சில சூழல்களில் இடம் பெயர்க்கலாம். ஒருபொருள்பன்மொழியக் குழுமத்திற்கு ஒரு பொருள்தான் உண்டு.

எ.கா.

(நூல், புத்தகம், புஸ்தகம்) "படிப்பதற்கு ஏற்றவகையில் அட்டைப்போட்டு இணைத்த அச்சடித்த தாள்களின் தொகுப்பு"

சொல்வலையில் ஒருபொருள்பன்மொழியக் குழுமங்கள் பொருண்மை உறவுகளால் இணைக்கப்பட்டுள்ளன. பெயர்களை ஒழுங்குபடுத்தப் பயன்படுத்தப்படும் மிக முக்கியமான உறவு உள்ளடங்கு உறவாகும் (hyponymy). இந்த உறவுதான் சொற்களைச் சொல்படிநிலை அமைப்பில் ஒழுங்குபடுத்துகின்றது.

8.6.4.4.1 சொற்களின் படிநிலை அமைப்பு

உள்ளடங்கு உறவுகள் பற்றிய செய்திகள் மரபு அகராதிகளில் விவரணைகளாகத் தரப்பட்டிருக்கும். எ.கா.

குயில்	இனிய குரலுடைய கரிய நிறப்பறவை.
பறவை	இரு கால்களும் அலகுகளும் உடைய, உடலில் இருபக்கங்களிலும் பறப்பதற்கு ஏற்றவகையில் சிறகுகளும் கொண்ட விலங்கினம்
விலங்கினம்	தானாகவே இயங்கும் உணர்வு உறுப்புகளும் செல்லுலோஸ் இல்லாத செல் சுவர்களும் உள்ள உயிரினம்.

உயிரினம்	உயிர் வாழ்கின்ற ஒன்று.

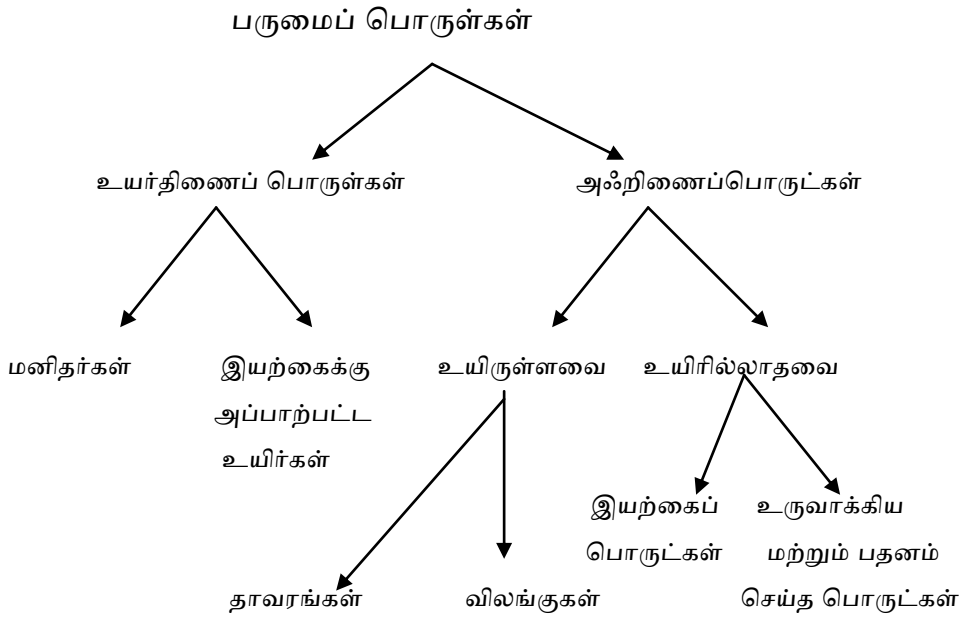
ஒவ்வொரு உள்ளடக்குச் சொல்லும் மிகப்பொதுவான உள்ளடக்குச் சொல்லுக்கு கொண்டு செல்லும். உள்ளடக்கு உறவைச் சொற்களுக்கு இடையே உள்ள உறவாக உருப்படத்தம் செய்ய இயலாது. உள்ளடக்கு உறவு அகராதிப்படுத்தப்பட்ட கருத்துக்களுக்கு இடையே உள்ள உறவாகும். ஒரு சொற்படிநிலை அமைப்பை உள்ளடக்கு உறவால் இணைக்கப்பட்ட ஒருபொருள்பன்மொழியக் குழுமங்களின் தொடர்ச்சியால் மீட்டுருவாக்கம் செய்யலாம்.

{குயில்}@->{பறவை} @ -> {விலங்கினம்}@ -> {உயிரினம்}

உள்ளடக்கு உறவுக்கு இணையான உள்ளடக்கு உறவையும் தரலாம். உள்ளடங்கு-உள்ளடக்கு உறவுகளால் ஒரு படிநிலை ஒழுங்குமுறை வெளிப்படும். இந்த வகையில் படும் படிநிலைகள் கணியியலார்களால் அறிவை உருப்படுத்தம் (knowledge representation) செய்யும் வழியாகப் பயன்படுத்தப்படுகின்றது.

8.6.4.4.2 தனித்தன்மையான தொடக்கிகள்

பெயர்கள் எல்லாவற்றையும் ஒரே படிநிலை அமைப்பில் தரவேண்டி படிநிலைக் கொள்கையை நீட்சி செய்யலாம். சொல்வலை பெயர்களைத் தனித்தன்மையான தொடக்கிகள் (unique beginners) அமையப் படிநிலை அமைப்புகளாகப் பகுத்துள்ளது. தனித்தன்மையான தொடக்கிகள் சொற்பொருண்மையியல் பொருண்மைக்கூறாய்வின் பொருண்மைக் கூறுகளுடன் ஓரளவுக்குப் பொருந்தும்.



மூலப்பொருண்மை அமைப்பைக் (ontological structure) காட்டும் வகையிலான சொற்களின் உருப்படுத்தம் சொற்களின் பொருண்மை மரபுரிமைச் செயற்பாங்கை (Lexical inheritance) வெளிக்கொணர்கின்றது.

எ.கா.

{ஊர்தி, வாகனம்}@ -> {நிலத்தில் ஓடும் ஊர்தி}@ -> {பொறியால் ஓடும் ஊர்தி}

@ -> நான்குசக்கர ஊர்தி}@ -> {பயண ஊர்தி} @ -> {பேருந்து}

யுரோ சொல்வலையின் பெயர் மூலக் கோப்புகளுக்கு 25 தனித்தன்மையான தொடக்கிகளின் பட்டியல் (List of 25 unique beginners for noun source files of Euro WordNet)

{act, activity}	{natural object}
{animal, fauna}	{natual phenomenon}
{artifact}	{person, human being}
{attribute}	{plant, flora}
{body}	{possession}
{cognition, knowledge}	{process}
{communication}	{quantity, amount}
{event, happening}	{relation}
{feeling, emotion}	{shape}
{food}	{state}
{group, grouping}	{substance}
{location}	{time}
{motivation, motive}	

8.6.4.4.3 வேறுபடுத்தும் பொருண்மைக் கூறுகள்

பெயர்களின் படிநிலை அமைப்பு உள்ளடங்கு-உள்ளடக்கு உறவால் உருவாக்கப்பட்டாலும் விளக்கங்கள் ஒரு கருத்துருவை மற்றொரு கருத்துருவிலிருந்து பிரிக்கும் பொருண்மைக் கூறுகளால் தரப்படுகின்றது.

எ.கா.

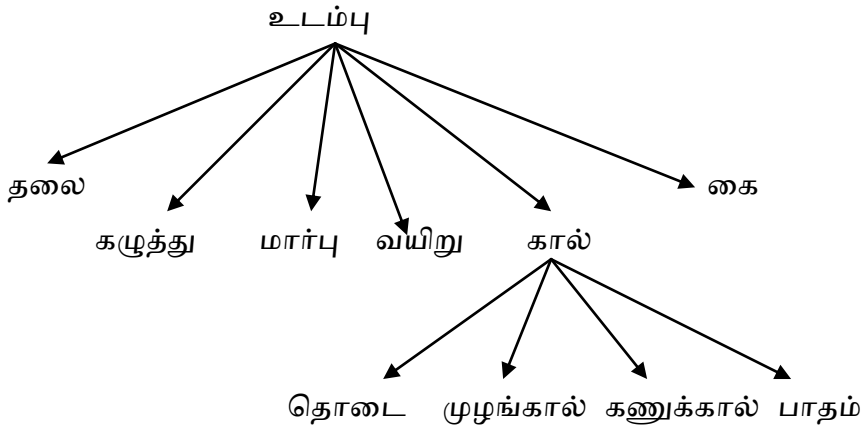
குயில் – ஒரு பாடுகின்ற சிறிய கரிய நிறப்பறவை

குயில் என்பதை மூன்றுவகைப் பட்ட வேறுபடுத்தும் பண்புக்கூறுகளுடன் தொடர்புபடுத்தலாம்.

1. அடைகள்: சிறிய, கரிய
2. பாகங்கள்: அலகு, சிறிய
3. செயல்பாடுகள்: பாடு, பற

8.6.4.4.4 சினை-முழு உறவு

சொற்களைப் படிநிலை அமைப்பில் தருவதில் சினை-முழு உறவு முக்கியப் பங்கு வகிக்கின்றது.



சினை-முழுஉறவு பின்வரும் துணை வகைகளை உள்ளடக்கும்:

1. முழுப் பொருளுக்கும் அதன் உறுப்புகளுக்கும் இடையிலான உறவை வெளிப்படுத்துவன (பாகம், எ.கா. கை – விரல்)
2. முழுமைக்கும் அதிலிருந்து பிரிந்த பகுதிக்கும் உள்ள உறவை வெளிப்படுத்துவன (பகுதி, எ.கா. துண்டு – உலோகம்)
3. இடங்களுக்கும் விரிந்த இடங்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (இடம், எ.கா. பாலைவனச்சோலை: பாலைவனம்)
4. குழுமத்திற்கும் அதன் அங்கத்தினர்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (எ.கா. மந்தை - ஆடு)
5. பொருள்களுக்கும் அது உருவாக்கப்பட்ட உருப்பொருள்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (ஆன, எ.கா. புத்தகம் – காகிதம்)

8.6.4.4.5. பெயர்ச் சொல்வலையின் பொருண்மை உறவுகளின் சுருக்கம்

பின்வரும் அட்டவணையில் பெயர் சொல்வலையில் கூறப்பட்டுள்ள பொருண்மை-சொல் உறவுகள் பட்டியலிடப்பட்டுள்ளன:

உறவுகள்	**துணை வகைகள்**	**எடுத்துக்காட்டுகள்**
ஒருபொருள் பன்மொழி		புத்தகம், நூல்
உள்ளடங்கு-உள்ளடக்கு		விலங்கு-பலூாட்டி
உள்ளடக்கு-உள்ளடங்கு		பசு-பாலூாட்டி
முழு-சினை	முழுமை-பாகங்கள்	மேசை-கால்
“	பாகம்-முழுமை	சக்கரம்-வண்டி
“	அங்கத்தினர்கள்-குழு	படைத்தலைவர்-படை
“	குழு-அங்கத்தினர்கள்	துறை-பேராசிரியர்
“	பகுதி-முழுமை	துளி-கண்ணீர்
“	இடம்-பரந்த இடம்	பாலைவனச்சோலை-பாலைவனம்
“	பொருள்-உருப்பொருள்	புத்தகம்-தாள்
ஈரிணை எதிர்நிலை	நிரலாக்கம் செய்யத்தக்கது	நல்லவன் – கெட்டவன்
“	துணை எதிர்நிலை	பகல் – இரவு
“	தனிப்பட்டவை	அஃறிணை– உயர்திணை
“	துருவ எதிரிநிலை	ஆண் – பெண்
“	பரஸ்பர சமூகப்பாத்திரங்கள்	மருத்துவர் – நோயாளி
“	சொந்தங்கள்	அம்மா – மகள்
“	கால உறவுகள்	காலை – மாலை
“	இட உறவுகள:செங்கோண	வடக்கு-கிழக்கு, மேற்கு
“	உறவுகள்	
“	இட உறவுகள்: நேர் எதிர்நிலை	வடக்கு –தெற்கு
“	பல்லிணை எதிர்நிலை: சங்கிலி	ஒன்று, இரண்டு, மூன்று
“	சுற்று	ஞாயிறு, திங்கள்,

		செவ்வாய், புதன், வியாழன், வெள்ளி, சனி

8.6.5. வினைச் சொல் வலை

ஒரு குறிப்பிட்ட செயலை அல்லது தெரிவைக் குறிக்கும் சொல் வினைகள் ஒரு மொழியின் சொல் மற்றும் தொடரியல் வகைப்பாட்டில் முக்கிய பங்கு வகிக்கின்றது. இதன் பயனிலை – பங்கெடுப்பாளர் அமைப்பு (predicate argument structure) இது வரும் வாக்கியத்தின் சாத்தியமான அமைப்புகளைத் தீர்மானிக்கின்றது.

எ.கா.

காற்று வீசுகின்றது.

அவன் சென்னையிலிருந்து தஞ்சாவூர் வந்தான்.

அவள் மாம்பழம் சாப்பிடுகிறாள்.

அவர் அவளிடமிருந்து ஒரு லட்ச ரூபாய்க்கு ஒரு கார் வாங்கினார்.

8.6.5.1 வினைகளைச் சொற்களங்களாகப் பிரித்தல்

யூரோ சொல்வலை வினைகளை 15 சொற்களங்களாகப் பிரிக்கின்றது. தமிழ்ச் சொல்வலை வினைகளைத் தற்காலிகமாக (இராஜேந்திரன், 2001) பௌதிக வினைகள் (*பெய், வீசு*), உடல்கூறு வினைகள் (*பிரசவி, வியர்*), புலனுணர்வு வினைகள் (*கச, குளிர்*), உணர்ச்சி வினைகள் (*கோபப்படு, வேதனையடை*), அறிவுசார் வினைகள் (*ஊகி, கணி*), கருத்துப்பாற்ற வினைகள் (*பேசு, இகழ்*), சலன வினைகள் (*நகர், நட*), தாக்க வினைகள் (*அடி, உடை*), உடைமைமாற்ற வினைகள் (*கொடு, விடு*), பல்கூட்ட வினைகள் (*சமையல் செய்*) எனப்பிரித்துள்ளது.

8.6.5.2 வினைகளுக்குத் தனித்தன்மையான தொடக்கிகள்

வினைச் சொற்களஞ்சியத்தைப் பொருண்மைக் களங்களாகப் பிரிப்பது வினைத் தரவுகளை ஒழுங்கு முறை படுத்துவதோடு சொற்களஞ்சியத்திலுள்ள எல்லா வினைகளுக்கும் ஒரு பொதுவான வேர் அல்லது தனித்தன்மையான தொடங்கிகள் இல்லாத குறையையும் நிவர்த்தி செய்கின்றது.

8.6.5.3 ஒருபொருள்பன்மொழியக்குழுமங்களாக வினைகள்

பெயர்களைப் போல வினைகளையும் ஒருபொருள் பன்மொழியக் குழுமங்களாகக் குழும இயலும். ஆனாலும் நாம் ஒருபொருள்பன்மொழி என்பது ஒன்றையொன்று இடம் பெயர்க்கும் சொற்கள் என்று வரையறை செய்தால் வினையில் ஒருபொருள்பன்மொழிகள் இல்லாமல் போகும் அல்லது மிக அரிதாகத் தான் இருக்கும்.

8.6.5.4 வினைகளின் பொருண்மைக் கூறாய்வு

வினைகளை அவற்றின் பொருண்மைப் பண்புக்கூறுகளால் விவரணை செய்ய இயலும். வினைகளைச் சிறு பொருண்மைக்கூறுகளாகப் பிரிக்க இயலும் தன்மை காரணமாக எளிய செயல்களால் கலவைத் தன்மையான செயல்களை விளக்க இயலும்.

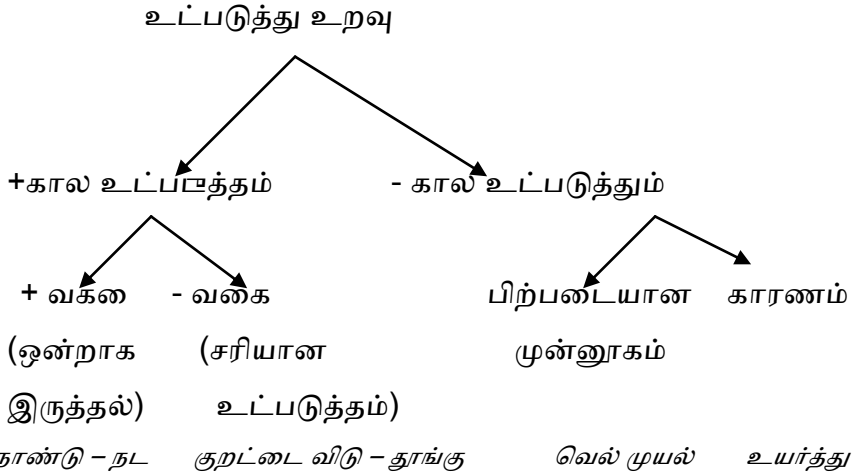
கொல் – சாகச் செய்தல்

எறி - ஒருவர் தன் கையிலிருந்து பொருளை விசையால் போகச் செய்தல்

ஒட்டு - ஒருவர் ஒன்றை ஓடச் செய்தல்

8.6.5.5 வினைகளுக்கிடையிலான சொல் மற்றும் பொருண்மை உறவுகள்

நாம் நான்கு வகையான உட்படுத்து உறவுகளை (entailment) வெளிப்படுத்தலாம்:



8.6.5.6 வினைகளுக்கிடையிலான பல்பொருள் ஒருமொழியம்

பெயர்களை விட வினைகளின் எண்ணிக்கை குறைந்தவை. வினைகளின் பல்பொருள் ஒருமொழியம் பெயர்களின் (polysemy) பல்பொருள் ஒருமொழியத்தைக் காட்டிலும் அதிகம். வினைகளின் பொருண்மை நெகழ்ச்சி அதன் பகுத்தாய்வைக் கடினப்படுத்துகின்றது. வினைகள் அவை எடுக்கும் பங்கெடுப்பவர் அமைப்பு அடிப்படையில் (argument structure) தங்கள் பொருளை மாற்றும். ஆனால் பெயர்களின் பொருண்மை வினைகளின் பொருண்மைகளைக் காட்டிலும் நிரந்தரமானது.

8.6.5.7 தொடரியல் பண்புகளும் பொருண்மை உறவுகளும்

தற்போது அகராதியில் சொற்களுக்கு அதன் தொடரியல் பண்புக்கூறுகளைத் தரும் போக்கு காணப்படுகின்றது. வினைகளைப் பொருண்மை உறவுகளாக மட்டும் பார்ப்பதும் தொடரியல் பற்றிய சில செய்திகளைத் தரும். வினையைப்பற்றி தெரிந்து கொள்ள வினையின் தொடரியல் பண்புகளைச் சொல் வலையில் தருவது குறித்த சாத்தியம் ஆயப்பட வேண்டும்.

8.6.5.8 வினைச் சொல்வலையின் பொருண்மை உறவுகளின் சுருக்கம்

பின்வரும் அட்டவணையில் வினைச் சொல்வலையில் கூறப்பட்டுள்ள பொருண்மை-சொல் உறவுகள் பட்டியலிடப்பட்டுள்ளன:

உறவுகள்	**விளக்கம்/துணை வகை**	**எடுத்துக்காட்டு**
ஒருபொருள் பன்மொழியம்	இடம்பெயர்க்க இயலும் செயல்கள்	தூங்கு – உறங்கு
பகுதி – முழுமை	உட்படும் – உட்படுத்தம் செயல்	பற – பிரயாணி
வகை உறவு	செயல் – துணை வகை	நட – நொண்டு
உட்படுத்து உறவு	செயல் – காரணச்செயல்	உயர் – உயர்த்து
“	செயல் – முன்னூகச் செயல்	
எதிர்நிலை	எதிர்நிலை	கூடு – குறை
	மறுதலை	வில் – வாங்கு
	திசை எதிர்நிலை	புறப்படு – வந்துசேர்

8.6.6 பெயரடை மற்றும் வினையடைச் சொல் வலை

சொல் வலை பெயரடைகளை இரு முக்கியமான வகுப்புகளாகப் பிரிக்கின்றது: வர்ணனைப் பெயரடைகள் (descriptive adjectives) மற்றும் தொடர்புப் பெயரடைகள் (relational adjectives). வர்ணனைப் பெயரடைகள் தலைப் பெயர்களுக்கு இருதுருவ அடைகளின் மதிப்பீட்டைத் (values of bipolar attributes) தருகின்றது; எனவே ஈரிணை எதிர்நிலைகள் (binary oppositions) (எதிர்மறை-antonymay) தருகின்றது; மற்றும் பொருண்மை ஒற்றுமையால் (similarity of meaning) ஒழுங்கமைக்கப்படுகின்றது.

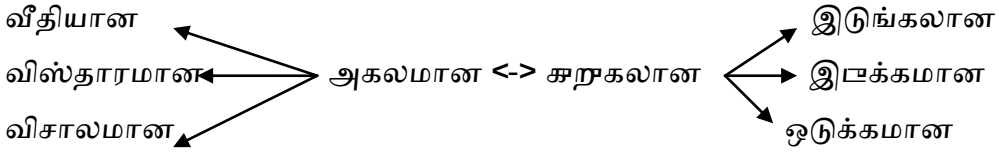
வர்ணனைப் பெயரடை (descriptive adjective): பெரிய, கனமான

தொடர்பப் பெயரடை (relational adjective): பொருளாதார, சகோதர

குறிப்பு அடைசெய்யும் பெயரடை (referemce modifying adjectives): பழைய, முன்னாள்

ஒரு வர்ணனை அடை ஒரு பெயருக்கு ஒரு மதிப்பு தரும். எடுத்துக்காட்டாக, *கனமான, இலேசான* என்பது *எடை* என்பதன் அடைமொழிகளாகும்; *தாழ்ந்த உயர்ந்த* என்பன *உயரம்* என்பதன் அடைமொழிகளாகும். எதிர்மறை (antonymy) தான் வர்ணனைப் பெயரடைகளில் அடிப்படையானதாகும். எதிர்மறை () தான் வர்ணனைப் பெயரடைகளில் அடிப்படையானதாகும். எதிர்மறை ஒருபொருள்பன்மொழியத்தைப் போன்று சொல்வடிவுகளுக்கு இடையிலான உறவாகும். எதிர்மறைப் பெயரடைகள் அடைமொழியின் எதிர்செய்யும் மதிப்புகளை வெளிப்படுத்துகின்றது (எ.கா. *கனமான, இலேசான* என்பன எடை அடைமொழியின் எதிர்

துருவங்களாகும்.) நேரடியான எதிர்மறைகள் இல்லாத பெயரடைகள் (எதிர்மறை ஆற்றல் குறைவாய் உள்ள பெயரடைகள்) எதிர்மறைப் பெயரடைகளுடன் "போல" உறவு கொண்டிருக்கும்:



உறவுகள்	**தொடர்புபடுத்தும் சொல்வகைப்பாடு**	**எடுத்துக்காட்டு**
எதிர்மறை (நிரலாக்கம் செய்ய இயலுவது)	பெயரடை/பெயரடை	அழகான/குரூரமான
எதிர்மறை (நிரலாக்கம் செய்ய இயலாதது)	பெயரடை/பெயரடை	உயிருள்ள/செத்த
ஆக்கவடிவு	பெயரடை/பெயர்	அழகான/அழகு
அடைமொழி	பெயர்/பெயரடை	வடிவம்/சின்ன
தொடர்பு	பெயரடை/பெயர்	பொருளாதார/பொருளாதாரம்
போல	பெயரடை/பெயரடை	பாரமான/கனமான

8.6.7 சொல்வலையை திட்டமிட்டு நடைமுறைப்படுத்தல்

சொல்வலையை நான்கு ஒழுங்கமைப்புகளாகப் பகுக்கலாம்.

1. சொல்வள ஒழுங்குமுறை (Lexial resource system)
2. ஒருங்கிணைக்கும் ஒழுங்கு முறை (Compiler system)
3. சேகரிப்பு ஒழுங்கு முறை (Storage system)
4. மீளப்பெறும் ஒழுங்கு முறை (Retrieval system)

அகராதியியலார் சொல்வலையின் மூலக் கோப்புகளை எழுதுகின்றனர். இது சொல்வள ஒழுங்குமுறையை அமைக்கும். ஒருங்கிணைக்கும் ஒழுங்குமுறை சொல்வள கோப்புகளை தரவுமைய வடிவாக உருவாக்குகின்றது. சேகரிப்பு ஒழுங்குமுறை சொல்வள ஒழுங்குமுறைக்கும் ஒருங்கிணைப்பு ஒழுங்குமுறைக்கும் இடையில் செயலாற்றுகின்றது. மீளப்பெறும் ஒழுங்குமுறை பின்வரும் செயல்பாடுகளை உள்ளடக்கும்:

1. வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை மின் தட்டச்சு அமைப்பாக மாற்றுகின்றது.
2. வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை தகவல் மீட்பு ஒழுங்கமைப்பாக மாற்றுகின்றது.
3. இணைய மீளப்பெறுதல்: வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை குறிதேடலுடன் இணைப்பது.

8.6.8.. முடிவுரை

சொல்வலை ஒரு சொல் தரவு மையம் (Lexical database). இதன் முக்கியமான பண்பு அதன் அர்த்தங்களின் வலைப்பின்னல். சொற்களுக்கு இடையே உள்ள ஒருப்பொருண்மையை அளந்து சொற்பொருள் மயக்கம் தீர்க்க சொல்வலை பயன்படுகின்றது. சொல்வலையைப் பயன்படுத்தி படிநிலை அமைப்பில் தரப்படுள்ள ஒரு சொற்களுக்கு இடையே உள்ள பொருண்மை உறவை ஆய்ந்து அச்சொற்களுக்கு இடையே உள்ள ஒற்றுமையை அளக்கலாம். பொருண்மை மயக்கம் தவிர மீட்புச்செயல்பாடுகளுக்கும் சொல்வலை பயன்படுகின்றது. தகவல் தளங்களிலும் தகவல் மையங்களிலும் இருந்து பயன்படுத்துபவரின் கேள்வியால் தேவையான தகவல்களைப் பெற இயலும். சொல்வலை இதற்குப் பெரிதும் உதவும். இயற்கை மொழி ஆய்வுக்கும் உருவாக்கத்திற்கும் இது பல வழிகளில் கைகொடுக்கும். சொல்வலையைப் பயன்படுத்தி இயந்திரமொழிபெயர்ப்பு செய்வதற்கான முயற்சிகளும் நடைபெற்று வருகின்றன. சொல்வலையில் உட்படுத்தப்பட்டுள்ள சொல் அறிவு அதைத் திறமையான ஒழுங்குமுறைகள் (expert systems), மொழிபெயர்ப்புத் துணைக்கருவிகள், தேடல் இயந்திரங்கள், கற்றல் ஒழுங்கமைப்புகள் மற்றும் தானியங்கு சுருக்கிகள் (antomatic summarizers) போன்றவற்றிற்குப் பொருத்தமானதாகச் செய்கின்றது. யூரோ சொல்வலை போலவே திராவிடமொழிகளின் சொல்வலை (Dravidian wordNet) உருவாக்குவதற்கான முயற்சிகள் தொடங்கப்பட்டு நடைமுறைப் படுத்தப்பட்டு வருகின்றன.

8.7. காட்சி மூலப்பொருண்மையியல் சொற்களஞ்சியம்

நைடா (1975) முன்வைத்த பொருளின் பொருண்மைப் பகுப்பாய்வின் கோட்பாடுகள், மரபு அடிப்பாடையிலான நிகண்டு மற்றும் அரிஸ்டாட்டிலியன் வகை மற்றும் இனங்கள் பற்றிய கொள்கைகளின் அடிப்படையில் 1990ஆம் ஆண்டில் தமிழிற்கான ஒரு சொற்களஞ்சியம் இராசேந்திரனால் உருவாக்கப்பட்டு 2001இல் வெளியிடப்பட்டது (இராஜேந்திரன், 2001). அதைத் தொடர்ந்து, தமிழிற்கான ஒரு மின்னணு சொற்களஞ்சியம் முயற்சிக்கப்பட்டு 2006ஆம்

ஆண்டில் தமிழ் மின்னணு சொற்களஞ்சியம் என்ற புத்தகம் வெளியிடப்பட்டது (இராஜேந்திரன் மற்றும் பாஸ்கரன், 2006). தமிழிற்கான சொல்வலை தயாரித்தல் (2001-2003; 2007-2015) திராவிடச் சொல்வலை உருவாக்கத்தின் ஒரு பகுதியாக இராசேந்திரனை முதன்மை ஆய்வாளராகக் கொண்டு அவரது மேற்பார்வையில் தமிழ்ப் பல்கலைக்கழகத்தின் இந்திய அரசின் மனிதவள மேம்பாட்டுத் துறை மற்றும் மின்னணு மற்றும் தகவல் தொழில்நுட்பத் துறையிலிருந்து பெறப்பட்ட நிதி உதவியுடன் கட்டமைக்கப்பட்டு ஒரு நிலையை அடைந்துள்ளது. இதன் பின்னர் இரண்டாயிரது பதினைந்து பதினாறுகளில் சென்னையிலுள்ள தமிழ் இணையக் கல்விக்கழக்தின் நிதி நல்கையில் இராஜேந்திரன் (இராஜேந்திரன், 2001) உருவாக்கிய சொற்களஞ்சியத்தில் தரப்பட்டுள்ள மூலப்பொருண்மையியல் அடிப்படையில் அமிர்தா பல்கலைக்கழகத்தில் கணினிசார் தொழில்நுட்டபம் மற்றும் வலைப்பின்னல் மையத்தில் (Centre for Computational Engineering and Networking) நடைபெற்ற "தமிழ் மொழி கற்பித்தல் மற்றும் கற்றலுக்கான கணினி கருவிகள்" ஆய்வுதிட்டத்தின் ஒரு பகுதியாகத் தமிழ் காட்சி மூலப்பொருண்மையியல் சொற்களஞ்சியம் (Rajendran 2016, 2018) உருவாக்கப்பட்டு அந்நிறுவனத்தின் இணையத்தில் பதிவேற்றம் செய்யப்பட்டுள்ளது. இச்சொற்களஞ்சியம் மேற்சொன்ன மின்சொற்களஞ்சியம், சொல்வலை இவற்றின் அடிப்படையில் அமையும்.

மெய்பொருள் மூலாய்வின் வரையறை விளக்கம்

மூலப்பொருண்மையியல் ஆய்வு ஒரு பகிர்ந்துகொள்ளப்பட்ட கருத்துருவாக்கத்தின் முறையான, வெளிப்படையான சிறப்பீடாகும். கருத்துவாக்கம் என்பது ஒரு நடத்தையின் தேவையான கருத்துருக்களின் கண்டுபிடிப்பால் உருவாகப்பட்ட ஒரு அருவ மாதிரியாகும். கருத்துருக்களும் அவற்றிற்கிடையேயுள்ள உறவுகளும் அவற்றின் பயன்பாட்டின் கட்டுப்பாடுகளும் வெளிப்படையாக வரையறை விளக்கம் செய்யப்பட்டுள்ளது. '"முறையானது" (formal) என்பது மூலப்பொருண்மையியல் ஆய்வு இயந்திரத்தால் படிக்க இயலுவது மற்றும் இயற்கை மொழியின் பயன்பாட்டை விலக்கி வைக்கின்றது. எடுத்துகாட்டாக, மருத்துவ பொருட்புலத்தில், கருத்துருக்கள் யாவும் நேய்களும் அறிகுறிகளும் ஆகும்; அவைகளுக்கிடையேயுள்ள உறவு இயல்பானது மற்றும் நோய் தானே காரணமாகாது என்பது கட்டுப்பாடாகும். மூலப்பொருண்மையியல் ஆய்வு ஒரு "பகிர்ந்துகொள்ளபட்ட கருத்துருவாக்கம்" என்பது ஒரு குழுவின் பயன்பாட்டிற்குரிய ஒருமித்த அறிவை உருப்படுத்தம் செய்வதை நோக்கமாகக் கொண்டது என்று கூறும். இலட்சிய அளவில் மூலப்பொருண்மையியல் ஆய்வு அதன் பயன்பாட்டிலிருந்து சுதந்திரமாகவும் உலகளவில் பகிர்ந்துகொள்ள இயலும் வழியிலும்

அறிவைக் கைகொள்ளும்; ஆனால் நடைமுறையில் வேறுபட்ட பணிகளும் பயன்பாடுகளும் மெய்ப்பொருள் மூலாய்வில் அறிவின் வேறுபட்ட உருப்படுத்தத்தை வேண்டும்.

மூலப்பொருண்மையியல் ஆய்வு சில வேளைகளில் வகைப்பாட்டியலுடன் குழப்பிக்கொள்ளப்படும். வகைப்பாட்டியல் என்பது ஒரு பொருட்புலத்திலுள்ள தரவின் வகைப்படுத்தலாகும். இரு முக்கியமான சூழல்களில் அவைகள் வேறுபடும்:

1. ஒரு மூலப்பொருண்மையியல் ஆய்வு கருத்துருக்களுக்கிடையே உறவுகளையும் கட்டுப்பாடுகளையும் உட்படுத்தியுள்ளதால் வளமான அகவயமான அமைப்பைக் கொண்டிருக்கும்.
2. மூலப்பொருண்மையியல் ஆய்வு பொருட்புலத்தின் அறிவைப் பற்றிய சில ஒருமித்த கருத்தை உருப்படுத்தம் செய்வதாக வலியுறுத்தும். இந்த ஒருமித்த கருத்து அறிவின் திட்டமிடப்பட்ட (இலக்கு) பயன்பாட்டாளர்களுக்கிடையில் ஆகும். எ.கா. சில நோய்களுக்கு வேண்டி ஒரு மருத்துவ மனையின் மூலப்பொருண்மையியல் ஆய்வை மருத்துவர்கள் பயன்படுத்துவது; வரலாற்றுக் கலையுடன் தொடர்புடைய மூலப்பொருண்மையியல் ஆய்வை கலைஞர்கள் பயன்படுத்துவது.

மூலப்பொருண்மையியல் ஆய்வுகள் அவை உருப்படுத்தம் செய்யும் அறிவைப் பற்றிய பொதுவான உடன்பாட்டின் ஒரு வடிவத்தை உருப்படுத்தம் செய்யவதை நோக்கமாகக் கொண்டுள்ளதால், அவைகள் பெருப்பாலும் வேறுபட்ட மக்களை உட்படுத்தி கூட்டுறவு செயல்முறையில் உருவாக்கப்படுகின்றன. மூலப்பொருண்மையியல் ஆய்வுகள் அவை கொண்டிருக்கும் கொள்கைகளின் பொதுமைத்தன்மையின் தரம் (degree) அடிப்படையில் பிரிக்கப்பட்டுள்ளது.

நைடாவின் சொற்றொகை வகைபாட்டு அடிப்படியில் மூலப்பொருண்மையியல் உருவாக்கம்

நைடாவின் பொருண்மையமைப்பு/ பொருட்புலப் பாகுபாடு அடிப்படையில் தமிழ்ச் (Nidat 1975a) சொற்றொகை மூலப்பொருண்மையியல் கோட்பாடு அடிப்படையில் காட்சி மூலப்பொருண்மையியல் சொற்களஞ்சியத்திற்காக அமைப்பாக்கம் செய்யப்பட்டுள்ளது.

நைடா (1975a) குறிப்புப்பொருளின் (Referential meaning) பொருளணுப் பகுப்பாய்வினை ஒரு மொழியின் சொற்றொகையின் பொருண்மை அமைப்பைக் காண்பதற்குப் பயன்படுத்துகிறார். சொல்லுக்கும் அதன் குறிப்புக்கும் (referent) உள்ள தொடர்புதான் குறிப்புப் பொருளாகும். குறிப்பு பொருள் இரண்டு காரணிகளை அடிப்படையாகக் கொண்டது. அவை (உணர்ச்சி பொருளுக்கு எதிரான) புலனறிவுப் பொருளும் (இலக்கணப் பொருளுக்கு எதிரான) குறிப்பு பொருளும் ஆகும். குறிப்புப் பொருள் ஒரு சொல்லின் கட்டமைந்த ஒரு கூட்டம் புலன்கூறுகள் (cognitive features)

அடங்கியதாகும். இப்புலன் கூறுகள் ஒரு சொல் மற்றொரு சொல்லிலிருந்து வேறுபடப் பயன்படுகின்றன. ஒரு பரப்பில் (semantic domain) அடங்கும் சொற்கள் தங்களுக்குள் வேறுபட உதவும் இம்முக்கியமான பொருண்மைக் கூறுகளை நாம் கண்டுபிடிக்க வேண்டும். ஒத்த பொருண்மைக் கூறுகளைக் கொண்டிருக்கும் குறிப்புப் பொருளை ஒன்றாக ஒரு பொருள் பரப்பில் அடக்கலாம். இப்பொருள் பரப்பின் தேவை (Relevancy), பருப்பம் (largeness), படிநிலை அமைப்பில் (காண்க பக்கம்:) அதன் செயற்பாட்டு நிலை மட்டம் (level of function) ஆகியன மொழியின் மொத்த பொருண்மைக் கட்டமைப்பைப் பொறுத்திருக்கின்றன.

ஒன்றுக்கொன்று தொடர்புடைய பொருள் தொகைகளின் பொருண்மைத் தொடர்புகளின் விளக்கம் இரண்டு வேறுபட்ட வகையான உறவுகளை எடுத்துக் கொள்ள வேண்டும். அவையாவன வெவ்வேறு சொற்களுக்கு இடையே உள்ள தொடர்புடைய பொருள்கள் மற்றும் ஒரு சொல்லின் தொடர்புடைய பொருள்களுக்கு இடையே உள்ள உறவு. ஒரு பொருள் பரப்பில் அடங்கும் வேறுபட்ட சொற்களின் பொருள்களை நான்கு வழிகளில் உறவுபடுத்தலாம். அவை உள்ளடங்கல் (inclusion), மேலுறல் (overlapping), துணைநிலைப்படுதல் (complementation), அண்மைப்படல் (contiguity), ஆகியன. உள்ளடங்கல் படிநிலைத் தொடரில் தங்களுக்குள் தொடர்புள்ள மற்றும் அவை அடங்கியுள்ள பொருளிலிருந்து குறைந்தது ஒரு வேறுபடுத்தும் பொருளணு கூடுதலாக உள்ள பொருள்கள் அடங்கிய தொகுதியை உருவாக்கும் (எ.கா. விலங்கு: பாலூட்டி). மேலுறல் ஒரு படிநிலை மட்டத்தில், மரபுப்படி ஒரு பொருள் பன்மொழிகள் என்று அறியப்படும் மற்றும் அவற்றின் மேலுறல் தொகுதியை உருவாக்கும் (எ.கா கொடு: அளி, காய்ச்சல்: சுகக்கேடு). துணைநிலை பொருள்களின் தொகுதிகள் மூன்று முக்கிய வழிகளில் தொடர்புடையன. அவை எதிர்நிலைச் சொற்கள் (opposites) (எ.கா. பெரிது: சிறிது) திருப்புப் பொருள் நிலைச் சொற்கள் (reversies) (எ.கா. கட்டு: அவிழ்) மறுதலைச் சொற்கள் (conversies) (எ.கா. வாங்குதல்: விற்றல்) ஆகியன. ஒரு படிமட்டத்தில் கூடுதல் பொருள் ஒப்புமை உள்ள அதே சமயம் குறைந்தது ஒரு முக்கிய பொருளணுவால் குறிப்பிடத்தக்க விதத்தில் வேறுபடும் பொருள்கள் ஒரு அண்மைப்பட்ட தொகுதியை உருவாக்கும் (எ.கா நாற்காலி: பெஞ்சு: ஸ்டூல்: சோபா).

ஒரு சொல்லின் வேறுபட்ட பொருள்களை நான்கு வழிகளில் தொடர்புபடுத்தலாம்: ஆக்கம் (derivation) மாற்றீடு (replacement) உருவகநீட்சி (figurative extension) மற்றும் புற எல்லைக் கோவை (peripheral clustering) அடிப்படைப் பொருளணுக்களை அதனின்றும் பெறப்பட்ட வேறு பொருள் பரப்பைச் சார்ந்த பொருளில் அடங்குவது ஆக்கமாகும். (எ.கா காதல்: காதலன்). அடிப்படைப் பொருளின் குறைந்தது ஒரு பொருளணுவால் பொருள் பரப்பில் பெரிய மாற்றம்

விளைவிக்காமல் இடமாற்றம் செய்வது மாற்றீடாகும் (எ.கா. ராஜா: ராணி). பெரிதும் வேறுபட்ட பொருள் பரப்புகளுல் அடங்கும். அடிப்படைப் பொருளுக்கும் உருவகப்பொருளுக்கும் உள்ள தொடர்பைச் சில துணைப் பொருளணுக்கள் (supplementaty components) அல்லது திரும்பி பொருள் கொள்ளப்பட்ட பொருளணுக்கள் (reinterpreted diagnostic components) வாயிலாக நிறுவது பொருளின் உருவக நீட்சியாகும் (எ.கா. நரி: தந்திரசாலி) மத்தியப் பொருளுடன் (central meaning) தொடர்புடைய அல்லது அத்துடன் சிலப் பொதுப் பொருளணுக்களைக் கொண்டு தங்களுக்குள் தொடர்புடைய புற எல்லைப் பொருள்களைக் கோவைப்படுத்துதல் புற எல்லைக் கோவையாகும். (எ.கா கெட்டியான தோல், கெட்டியான பால் இவற்றில் வரும் கெட்டி).

நைடா (1975b:152) பின்வருமாறு கூறுகிறார்: பொருள் உணர்த்தும் பலவகை சொற்சேர்க்கையின் ஆய்வு நம்மை நான்கு அடிப்படை பொருண்மைக் குறியீடுகளின் வகுப்புகளை இனங்கண்டுகொள்ளத் தூண்டும். இவ்வகுப்புகளில் அடங்குபவை: 1. குதிரை, மரம், இலை, சூரியன், மலை, போன்ற 'பருப்பொருள்கள்' (objects), 2. ஆடுதல், நடத்தல், வெட்டுதல், தள்ளுதல், போதல், இருத்தல், புரள்தல், போன்ற 'நிகழ்வுகள்' (events), 3. அ) சிவப்பு, நீலம், சின்னது, பல, ஒன்று, இரண்டு, (பருப்பொருள்களின் அருவங்கள்), ஆ) விரைவு, மெல்ல, ஒருதடவை, இரண்டு தடவை, அடிக்கடி, (நிகழ்வுகளின் அருவங்கள்), இ) கடினம் (எ.கா கடினமான பாறை, கடினமான வேலை) மிருது (எ.கா மிருதுவான கோட்டு, மிருதுவான வேலை) (பருப்பொருள்கள் அல்லது நிகழ்வுகள் என்பதன் அருவங்கள்) அல்லது ஈ) மிக (எ.கா மிகப்பல) ரொம்ப (எ.கா ரொம்ப சின்ன) வெகு (எ.கா வெகு விரைவு) (அருவங்களின் அருவங்கள்) இவற்றைத் தொடர்புபடுத்த உதவும் 'தொடர்பான்கள்' (rationals), எ.கா அ) பருப்பொருளுடன் பருப்பொருள்: ஜானும் பில்லும் , வீட்டுள்ள மனிதர்கள், மேஜை மேலுள்ள புத்தகம்: ஆ) பருப்பொருளுடன் நிகழ்வுகள்: பறக்கும் மனிதர்கள், வேட்டை நாய்களின் குரைப்பு ; இ) நிகழ்வுடன் நிகழ்வு: ஓட்டமும் குதிரையும், நீச்சலில் வெற்றி; ஈ) பருப்பொருளுடன் அருவம்; பொருளின் அழகு, ரோஜாவின் நிறம்; உ) அருவத்துடன் அருவம்: மோசமும் சோம்பலும், பரிசுத்த அழகு. இந்த நான்கு அடிப்படைப் பொருண்மை வகுப்புகளும் எல்லா மொழிகளிலும் காணப்படுவதுடன் எல்லா வகையான சொல்லுருக்களுக்கு இடையே உள்ள பொருள் தொடர்பை ஆய்வதற்கும் அடிப்படையாக அமையும் என்றும் தோன்றுகின்றது.

நைடா கிரேக்கப் புதிய வேதாகமத்திற்கு ஒரு புதிய அகராதி உருவாக்க முனைந்து அதை நிறைவேற்றியுள்ளார். இது பொருள்களைச் சொல் அலகுகளின் அகர வரிசை நிரலில் அல்லாமல் பொருள் பரப்பின் அடிப்படையில் முறைப்படுத்தும். வெவ்வேறு பொருண்மை அலகுகளின் தொடர்பான அர்த்தங்களின் அடிப்படையில் பொருள் பரப்புகள் பகுக்கப்பட்டுள்ளன.

தேவையான பொருணுக்களைக் காணுவதில் ஒரு பொருள் பரப்புகள் பகுக்கப்பட்டுள்ளன. தேவையான பொருள்களைக் காணுவதில் ஒரு பொருள் பரப்பின் குறிப்பு நுட்பம் பொதுவாக எந்த ஒரு பொருள் தொகுதியும் தம்முள் அடங்கியுள்ள பொதுவான பொருளணுக்களின் எண்ணிக்கையோடு தகவுப்பொருத்தம் கொண்டிருக்கும். இந்த அகராதி முக்கியமாக மொழிகளுக்கு இடையே உள்ள பொருளணு அடிப்படையிலான வேறுபாடுகளை அறிந்த பைபிள் மொழிபெயர்ப்பாளர்களுக்காகத் திட்டமிடப்பட்டதாகும். நாம் முன்னர் கண்ட சொல் உறவுகளின் அடிப்படையில் பொருளணு ஆய்வைப் பயன்படுத்தி அவர் கிரேக்க மொழியின் சொற் கட்டமைப்பைத் தர முயன்றுள்ளார். கீழே தரப்பட்டுள்ளது நைடா (1975a:176-186) தந்துள்ள வகைப்பாட்டின் சுருக்கமாகும்.

1. பருப்பொருள்

எ. விலங்குகளல்லாதவை

1. இயற்கையானவை

எ. பூகோளம் சம்பந்தப்பட்டவை

பி. இயற்கை சாதனங்கள்

சி. தாவரங்கள் மற்றும் தாவரப்பொருட்கள்

2. உருவாக்கிய அல்லது கட்டப்பட்ட பருப்பொருள்கள்

எ. கலைப்பொருள்

பி. பதம் செய்யப்பட்ட பொருள்கள்: உணவுகள் மற்றும் வாசனை பொருள்கள்

சி. கட்டிடங்கள்

பி. விலங்குப்பொருள்கள்

1. விலங்குகள், பறவைகள், பூச்சிகள்
2. மனிதர்கள்
3. இயற்கைக்கு அப்பாற்பட்ட சக்திகள் அல்லது உயிர்கள்

நிகழ்வுகள்

பௌதீகம் தொடர்பானவை, உடற்கூறு தொடர்பானவை, உணர்வு தொடர்பானவை, உணர்ச்சி தொடர்பானவை, அறிவு தொடர்பானவை, கருத்துப்பரிமாற்றம், சேர்க்கை, கட்டுப்பாடு, சலனம், தாக்கம், கைமாற்றம், தொடர்ச்சியான சலனங்கள் அல்லது செயல்கள் உள்ளடங்கிய பல்கூட்டச் செயல்கள்.

அருவங்கள்

காலம், தூரம், பருமை, விரைவு, சூடு, நிறம், எண், நிலை, மதப்பண்பு, கவர்ச்சி, வயது, உண்மை-பொய், நல்லது-கெட்டது, பலம், ஆரோக்கியம் நிலை.

தொடர்பான்கள்

இடம் தொடர்பானவை, காலம் தொடர்பானவை, கட்டல், தர்க்கம், தொடர்பானவை.

மேற்சொன்ன வகைப்பாடு குறிப்புப் பொருள்களை அடிப்படையாகக் கொண்டது. சொற்பரப்பு வகுப்புகளுக்கும் இலக்கண வகுப்புகளுக்கும் ஒன்றுக்கொன்று நிகரான பொருத்தம் இல்லை. இருப்பினும் அவை தம்மில் சில இசைவுகள் இருக்கத்தான் செய்கின்றன. ஏனென்றால் ஏதோ ஆழ்நிலை அமைப்பு மட்டத்தில் பருப்பொருட்கள் பெயர்களாலும் நிகழ்வுகள் வினைகளாலும் அருவங்கள் அடைகளாலும் தொடர்பான்கள் பல வேறுபட்ட கூறுகளாலும் (இடைச் சொற்கள், வேற்றுமை ஒட்டுகள், சொல் நிரல் போன்றவை) குறிக்கப்படுகின்றன. மேற்கண்ட வகைப்பாடு நிச்சயமாக ஒரு தெளிவான தர்க்க நோக்கால் அமைந்ததாகும். இருப்பினும் இதே சிந்தையை அடிப்படையாகக் கொண்டு சிறிது வேறுபட்ட ஒரு வகைப்பாட்டைச் செய்ய இயலும்.

நைடா உலகப்பொருண்மை வகைப்பாட்டை அதிகமான மாற்றங்கள் செய்யாமல் தமிழுக்கு ஏற்றுக் கொள்ளலாம் என்றாலும் இவ்வாறு செய்வதிலும் பல சிக்கல்கள் வரத்தான் செய்கின்றன. நைடாவே தன் வகைப்பாட்டின் குறைப்பாடுகள் பற்றிக் குறிப்பிடுகிறார் (Nida, 1975:187): "இந்தச் சொற்பரப்புகளின் பட்டியல் தற்காலிகமானதாகும். அகராதிச் சொற்களஞ்சியத்தை (Dictionary-thesarus) உருவாக்கும் குழுக்கள் பதிப்புவேலையைத் தொடர்ந்து கொண்டிருப்பதால் இந்த வகைப்பாடு மாற்றத்திற்கு உரியதாகும். மேலும், இதில் கொடுக்கப்பட்டுள்ள பட்டியல் முழுமையானதல்ல. ஏனென்றால் இது ஒரு சில சொற்பரப்புகளைத்தான் பட்டியல் படுத்தியுள்ளது. பெரும்பாலும் ஒரு சில மாதிரிக்கான பொருள்கள் தாம் கொடுக்கப்பட்டுள்ளன." (நைடாவின் கிரேக்க அகராதிச் சொற்களஞ்சியம் வெளிவந்துவிட்டது.) நைடாவின் பொருண்மைச் சட்டகத்தில் பொருந்தும் படி தமிழ்ச் சொற்களை வகைப்பாடு செய்வது சில இடங்களில் கடினமாக அமைகிறது. ஒரு குறிப்பிட்ட சொல் எந்தப் பொருள் பரப்பைச் சார்ந்தது என்பதை முடிவு செய்வதில் அடிக்கடி தெளிவின்மை ஏற்படுகிறது. நைடாவே பின்வருவனவற்றை முக்கியமான சிக்கல்களாக காட்டுகிறார். 1. சொல்லாக்கப் பொருளுக்கும் அதன் அடிப்படைக்கும் உள்ள உறவு, 2. சொற்திரிபு வேற்றுமைகளைக் கையாளுதல், 3. ஒரு சொல்லின் வேறுபட்ட பொருள்கள் வேறுபட்ட மட்டங்களில் வருதல்,4. சொற்பரப்புகளுக்குப் பொதுப்பொருள் வருதல் அல்லது வராமை, 5. படிநிலை வகைப்பாட்டின் கட்டமைப்பு, 6. சொற்பரப்புகளுக்கு இடையே உள்ள எல்லைகள் இச்சிக்கல்களை மனதில் கொண்டு சொற்களஞ்சியப் பொருட்புல வகைப்பாடு செய்யப்படவேண்டும்.

நடாவின் சொற்றொகைப் பாகுபாடுஅடிப்படையில் தமிழ்ச்சொற்றொகை பகுக்கப்பட்டு சொல்வலையில் விளக்கப்பட்டுள்டுள்ள சொல்சார் உறவுகள் மற்றும் பொருண்மை உறவுகள் அடிப்படையில் உறவுபடுத்தப்பட்டுள்ளன. இதுதான் இச்சொற்களஞ்சியத்தின் தரவுதளமாகும்.

காட்சி மூலப்பொருண்மையியல் சொற்களஞ்சிய உருவாக்கம்

ஒரு பயனரின் வினவலுக்காக கொடுக்கப்பட்ட மிகப்பெரிய தரவுத்தளத்திலிருந்து மூலப்பொருண்மையியல் கிளையமைப்பை உருவாக்குவது என்பது ஒரு புத்திசாலித்தனமான தகவல் மீட்டெடுப்பு மற்றும் காட்சிப்படுத்தல் ஒழுங்கமைப்பாகும். இந்த வகைக் கருவி தரவு உருப்பதுத்தம் மற்றும் அறிவுப் பிரித்தெடுத்தல் (பெறுதல்) செயல்முறைக்கு ஒரு சக்திவாய்ந்த வழியை வழங்குகிறது. பயனர் கணினியில் தேடவேண்டிய வார்த்தையை உள்ளீடு செய்து தரவுத்தளத்திலிருந்து தகவல்களைப் பெறலாம். தகவல் ஒரு காட்சி உருப்பத்தமாகக் கொடுக்கப்பட்டுள்ளது. இந்த நோக்கத்திற்காக நன்கு கட்டமைக்கப்பட்ட தரவு தளம் தேவை. பின்வரும் அட்டவணை சொற்கள் அல்லது சொற்பொருள் இவைகளுக்கு இடையில் நிறுவப்பட்ட பல்வேறு வகையான அர்த்த உறவுகளையும் மூலப்பொருண்மையியல் கிளை அமைப்பு உருவாக்குவதற்குத் தரவுத்தளத்தில் அவற்றைக் குறிக்கும் தனித்துவமான வழியை விளக்குகிறது. தற்போதைய காட்சிச் சொற்களஞ்ச்யத்தில் பின்வரும் எழு உறவுகளே நிறுவப்பட்டுள்ளன. இவ் உறவுகளின் எண்ணிக்கை எதிர்காலத்தில் அதிகரிக்கப்படும்.

பொருண்மை மற்றும் சொல்சார் உறவுகள்	**விளக்கமும் எடுத்துக்காட்டும்**	**உறவை நிறுவும் குறிப்பான்கள்**
இணைமொழியம் (ஒருபொருள்பன்மொழியம்)	ஒரே பொருண்மையைக் குறிக்கும் சொற்களுக்கு இடையே உள்ள உறவு. எ.கா. புத்தகம், நூல் என்பவைகளுக்கு இடையே உள்ள உறவு.	**< >**
வகைமொழியம்	உள்ளடங்கு மொழிக்கும் உள்ளடக்கு மொழிக்கும் இடையே உள்ள உறவு. நாய், பூனை போன்ற உள்ளடங்கு மொழிகளுக்கும் விலங்கு என்ற உள்ளடக்கு மொழிக்கும் இடையில் உள்ள உறவு.	{ }

பகுதிமொழியம்	சினைமொழிக்கும் முழுமொழிக்கும் உள்ள உறவு. கை,கால், தலை என்ற சினைமொழிகளுக்கும் உடல் என்ற சினை மொழிக்கும் உள்ள உறவு.	/ |
எதிர்மொழியம்	எதிர்மொழியம்	% *
பால்மொழியம்	பலர்பால் ஒருமை வடிவிற்கும் ஆண்பால் ஒன்றன்பால் மற்றும் பொண்பால் ஒன்றன்பால் வடிவிற்கும் உள்ள தொடர்பு. எடுத்துக்காட்டாக வேலைக்காரர் என்பதற்கும் வேலைக்காரன், வேலைக்காரி என்பதற்கும் உள்ள உறவு	$ #
வினைப்பெயராக்கமொழியம்	வினைக்கும் அதிலிருந்து உருவாக்கப்பட்ட பெயர்ச்சொற்களுக்கும் இடையிலுள்ள ஆக்க உறவு வினைப்பெயராக்கமொழிய உறவு என்று அழைக்கப்படும். எடுத்துக்காட்டாக படி என்ற வினைக்கும் படிப்பு, படித்தம், படித்தல் என்ற வினைப்பெயர்களுக்கும் உள்ள உறவு.	~ `
அடைமொழியம்	பெயருக்கும் அதிலிருந்து ஆக்கப்பட்ட பெயரடைகளுக்கும் வினையடைகளுக்கும் இடையே	^ !

	உள்ள உறவு. எத்துக்காட்டாக *அழகு* என்ற பெயருக்கும் அதிலிருந்து உருவாக்கப்பட்ட *அழகான* என்ற பெயரடை மற்றும் *அழகாக* என்ற வினையடை இடையே உள்ள உறவு.	

ஒருபொருள் பன்மொழியம்

ஒரே பொருளைக் கொண்ட அல்லது பொருளின் ஒரே பொருண்மைக்கூறுகளைப் பகிர்ந்து கொள்ளும் சொற்களுக்கு இடையே உள்ள உறவு ஒருபொருள் பன்மொழியம் எனப்படும்; இவ் உறவைக் காட்டும் சொற்கள் ஒழுபொருள் பன்மொழிகள் எனப்படும். ஒத்த பொருள் என்பது சம்பந்தப்பட்ட சொற்கள் அர்த்த வெளிப்படுத்தத்தில் ஒரே மாதிரியாக இருக்க வேண்டும் என்று அர்த்தமல்ல; கருத்தப்பட்ட சூழல்களில் வாக்கியத்தின் அர்த்தத்தில் எந்த வித்தியாசமும் இல்லாமல் ஒன்றுக்கொன்று இடம்பெயர்க்க இயலும் சொற்கள் ஒருபொருள் பன்மொழிகளாகும். உதாரணமாக, சொற்களை எடுத்துக் கொள்ளுங்கள். எடுத்துக்காட்டாக *நான் புத்தகம் படித்தேன்* என்பதும் *நான் நூல் படித்தேன்* என்பதும் ஒரே பொருண்மையைத் தருவதால் *புத்தகம்* என்பதும் *நூல்* என்பதும் ஒருபொருள் பன்மொழிகளாகும். ஒருபொருள் பன்மொழியச்சொற்கள் தரவுத்தளத்தில் '**< >**' என்ற அடைப்புக் குறிக்குள் தரப்பட்டிருக்கும். பின்வருவது ஒரு எடுத்துக்காட்டாகும்.

> சூரியன் **<** ஞாயிறு, ஆதவன், பகலவன், திவாகரன், ஆதித்தன், கதிரவன், கதிரோன், கிரணன், செங்கதிரோன், செங்கதிர், வெங்கதிரோன், வெங்கதிர், வெஞ்சுடர் ஆயிரங்கதிரோன், ஆயிரங்கிரணன், உலகநேந்திரன், உலகப்பாந்தவன் **>**; { எழுஞாயிறு, ஏறுஞாயிறு; சாய்ஞாயிறு, இறங்குஞாயிறு; இளஞ்சூரியன், **<** பாலசூரியன் **>**; }.

உள்ளடங்குமொழியம்-உள்ளடக்குமொழியம் (Hyponymy and Hypernymy)

உள்ளடங்கு மொழி என்பது ஒருசொல் மற்றொருசொல்லின் கீழ் பொருண்மை அடிப்படையில் உட்படுத்தப்படுவதாகும். உட்படும் சொல் உள்ளடங்கு மொழியம் என்றும் உட்படுத்தும் சொல் உள்ளடக்கு மொழியம் என்றும் அழைக்கப்படும். எடுத்துக்காட்டாக நாய், பூனை என்ற உள்ள்டங்குமொழிகள் விலங்கு என்ற உள்ள்டக்கு மொழியின் கீழ் பொருண்மை

அடிப்படையில் உட்படுத்தப்படும். நாய், பூனை என்பன விலங்கு என்பதன் இணை உள்ளடங்கு மொழிகளாகும். தரவுதளத்தில் இணை உள்ளடங்கு மொழிகள் உள்ளடக்கு மொழிகளுக்கீழ் மீசை அடைபுக்குறிகளுக்குள் தரப்பட்டுள்ளன. "{ }". பின்வருவது எடுத்துக்காட்டாகும்.

சப்தமண்டலம்: {வாயுமண்டலம்; வருணமண்டலம்; சந்திரமண்டலம்; சூரியமண்டலம்; நட்சத்திரமண்டலம்; அக்கினிமண்டலம்; திரிசங்குமண்டலம்; }

நவகிரகம்: {சூரியன்; சந்திரன்; செவ்வாய்; புதன்; வியாழன்; சுக்கிரன்; சனி, ராகு, கேது}

சினைமொழியம்-முழுமொழியம் (Meronymy-Holonymy)

சில சொற்கள் அவை மற்றொரு சொல்லின் உறுப்பாக வரும் சொல் என அகராதியில் விளக்கப்பட்டுள்ளதைக் காணலாம். உறுப்பாய் வரும் சொற்கள் சினைமொழி என்றும் முழுமையாய் (உடலாய்) வரும் சொல் முழுமொழி என்றும் அழைக்கப்படும். எடுத்துக்காட்டாக *கை, கால்* என்ற சினைமொழிகள் உடல் என்ற முழுமொழியுடன் உறவு படுத்தப்பட்டிருக்கும். தரவுதளத்தில் சினைமொழிகள் அவற்றின் முழுமொழிக்குக் கீழ் "/ |" என்ற குறியீடுகளுக்குள் தரப்பட்டிருக்கும். பின்வருவது எடுத்துக்காட்டாகும்.

உடல்: /தலை; கழுத்து; கை,கால்; மார்பு; வயிறு; புருடம்|

காற்றுமண்டலம்: /வெளிமண்டலம்; வெப்பமண்டலம், உஷ்ணமண்டலம்; படலம்: அயனிப்படலம்; ஓசோன்படலம்; |

எதிர்மொழியம்

இரண்டுசொற்கள் எதிரிடையான பொருள்கள் தருவதாகக் கருதப்படுமானால் அவைகளுக்கு இடையே உள்ள உறவு எதிர்மொழிய உறவு எனப்படும். அவைகள் ஒன்றுக்கொன்று எதிர்மொழிகளாக அமையும். *பெரிய* என்பது *சிறிய* என்பதன் எதிர்மொழியாகும் அல்லது *பெரிய, சிறிய* என்பன எதிர்மொழிகளாகும். எதிர்மொழிகள் தரவுதளத்தில் "% *" என்ற குறியீடுகளால் இணைக்கப்பட்டிருக்கும். பின்வருவது எடுத்துக்காட்டாகும்.

இருட்டான %வெளிச்சமான*; குட்டியான %நீளமான*; சூடான %குளிர்ச்சியான*

பால்மொழியம்

சில ஆண்பாலுக்கும் பெண்பாலுக்கும் அடையாளப் படுத்தப்பட்ட சில சொற்களுக்கும் அவற்றின் பலர்பாலைக் குறிக்கும் சொற்களுக்கும் இடையே உள்ள உறவு பால்மொழியம் என இங்கு குறிப்பிடப் படுகின்றது. எடுத்துக்காட்டாக வேலைக்காரர் என்பதற்கும் வேலைக்காரன், வேலைக்காரி என்பதற்கும் உள்ள உறவு

பால்மொழியம் எனப்படும். தரவுதளத்தில் பலர்பால் சொல்லுக்கும் பால் அடையாளப்படுத்தப்பட்ட சொற்களுக்கும் இடையே உள்ள உறவு பால் உறவு எனப்படும். இவைகளுக்கு இடையே உள்ள உறவு தரவுத்தளத்தில் பின்வரும் குறியீட்டால் உறவுபடுத்தப்படும்: "$ #". பின்வருவது எடுத்துக்காடாகும்.

இளைஞர்கள்: {இளைஞர் $ இளைஞன் #, வாலிபர் $ வாலிபன் #, வாலிபப்பையன், இளவல், இளந்தாரி, இளைஞோர் $ இளைஞோன் #, காளை, விடலை, இளவட்டம், வயசுப்பையன், வயசுப்பிள்ளை, வயசுப்பிள்ளையாண்டன், இளம்வயதினர், $இளம்வயதினன்#, இளம்பருவத்தார் $ இளம்பருவத்தான்#, இளம்பிராயத்தார் $ இளம்பியாயத்தான்#, பதின்பருவத்தினர் $ பதின்பருவத்தினன் #, பதின்வயதினர் $ பதின்வயதினன் #, வயதுவந்தோர், <வயதுவந்தவர்> $ வயதுவந்தவன் #, இளந்தை, இளந்தாரி, வாலியன், பாலியன்; இளரத்தம்> }

வினைப்பெயராக்கமொழியம்

வினைக்கும் அதிலிருந்து உருவாக்கப்பட்ட பெயர்ச்சொற்களுக்கும் இடையிலுள்ள ஆக்க உறவு வினைப்பெயராக்கமொழிய உறவு என்று அழைக்கப்படும். எடுத்துக்காட்டாக *படி* என்ற வினைக்கும் *படிப்பு, படித்தம், படித்தல்* என்ற வினைப்பெயர்களுக்கும் உள்ள உறவு வினைப்பெயராக்கமொழியம் எனப்படும். வினைக்கும் அதனுடன் தொடர்புள்ள வினைப் பெயர்களும் தரவுத்தளத்தில் பின்வரும் குறியீடுகளால் உறவுப்படுத்தப்படும்: "**~ `".**

நிகழ்வுகள்குறித்தவை: { நிகழ்; ~ நிகழ்தல்; நிகழ்ச்சி; நிகழ்வு; ` < சம்பவி; ~ சம்பவித்தல்; சம்பவம்; ` விளை; > ~ விளைதல்; விளைவு; ` >நேர்; ~ நேர்தல்; ` பலி; ~ பலித்தல்; ` ஈடேறு; ~ ஈடேறுதல்; ஈடேற்றம்; ` நிலவு; ~ நிலவுதல்; ` }

அடைமொழியம்

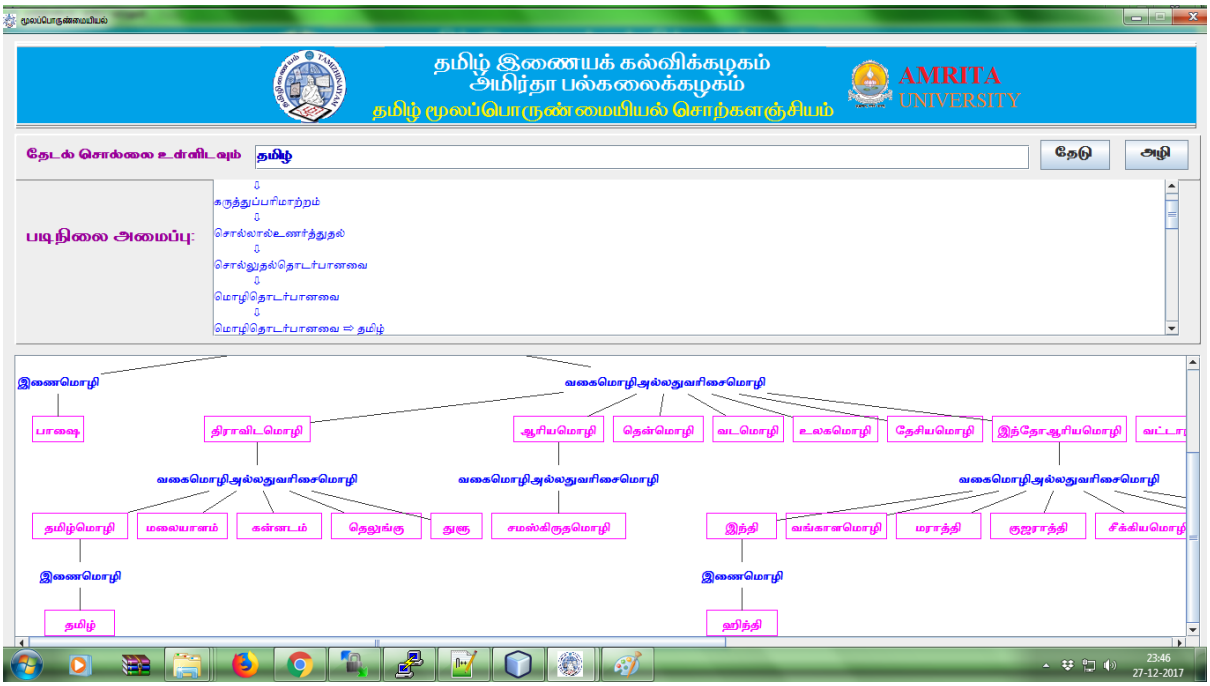
பெயருக்கும் அதிலிருந்து ஆக்கப்பட்ட பெயரடைகளுக்கும் வினையடைகளுக்கும் இடையே உள்ள உறவு அடைமொழியம் எனப்படும். எத்துக்காட்டாக *அழகு* என்ற பெயருக்கும் அதிலிருந்து உருவாக்கப்பட்ட *அழகான* என்ற பெயரடை மற்றும் *அழகாக* என்ற வினையடை இடையே உள்ள உறவு அடைமொழியம் ஆகும். பெயர்கள்

அதிலிருந்து உருவாக்கப்பட்ட அடைமொழிகளுடன் பின்வரும் குறியீடுகளால் உறவுபடுத்தப்படும்: "^ !". பின்வருவனது எடுத்துக்காட்டாகும்:

நீண்டகாலம்தொடர்பானவை: { நீண்டகாலம்; ^ நீண்டகாலமாக !: நீடுழி, < நெடுங்காலம் ^ நெடுங்காலமாக; ! ரெம்பகாலம்; ^ ரெம்பகாலமாக; ! நீண்டநேரம், < நெடுநேரம், ரெம்பநேரம், நிறையநேரம்; > நீண்டநாள், < நெடுநாள்; ரெம்பநாள், நிறையநாள் >; பல்லாண்டு;நூறாண்டு, < நூற்றாண்டு > }

பயனர் இடைமுகம்

பின்வரும் இடைமுகத்தில் பயனர்கள் வேண்டிய சொல்லை உள்ளீடு செய்து மூலப்பொருண்மையியல் சார் கிளைக் காட்சியைப் பெறலாம். பின்வரும் இடைமுக எடுத்துக்காட்டில் *தமிழ்* என்ற சொல் உள்ளீடு செய்யப்பட்டு அது தொடர்பான தகவல்கள் கிளைப்படமாகத் தரப்பட்டுள்ளன.



8.8.சுருக்க உரை

இவ்வியலில் சொல்சார் மூலவளங்களை உருவாக்குவதற்கான கணினிச் செயல்பாடுகளும் மொழியியல் செயல்பாடுகளும் விரிவாகத் தரப்பட்டுள்ளன. குறிப்பாக மின்சொற்களஞ்சிய உருவாக்கம், இயந்திரம் படிக்கவியலும் அகராதிகள் உருவாக்கம், சொல்வலை, காட்சி மூலப்பொருண்மையியல் சொற்களஞ்சியம் என்பனவற்றின் உருவாக்கம் விரிவாக விளக்கப்பட்டுள்ளது.

Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

இயல் 9

சொற்பொருண்மை மயக்கநீக்கம்

9.1.முன்னுரை

மக்கள் பொருண்மை மயக்கம் காட்டுகின்றார்கள் என்று கூறுவது அவர்கள் எவ்வாறு மொழியைப் பயன்படுத்துகின்றனர் என்பதைக் குறிக்கும் என்றாலும், பொருண்மை மயக்கம் என்பது மொழி வெளிப்பாடுகளின் பண்பாகும். ஒரு சொல்லோ தொடரோ வாக்கியமோ ஒன்றிற்கும் மேற்பட்ட பொருண்மைகளைக் கொண்டிருந்தால் அவை பொருண்மை மயக்கம் கொண்டுள்ளன என்று பொருள்படும். வெளிப்படையாக இந்த வரையறை விளக்கம் எந்தப் பொருள்கள் அல்லது ஒரு வெளிப்பாடு ஒன்றைக் கொண்டிருக்கும் என்பது என்ன என்று கூறவில்லை. ஒரு குறிப்பிட்ட மொழிக்கு இந்தத் தகவல் இலக்கணத்தால் தரப்படும்; இது முறையாக வடிவங்களைப் பொருண்மைகளுடன் பொருத்தும் மற்றும் பொருண்மை மயக்கம் உள்ள வடிவங்களை ஒன்றிற்கும் மேற்பட்ட பொருண்மைகளுடன் பொருத்தும்.

சொற்பொருண்மை மயக்கம் தொடரிலும் வாக்கியத்திலும் பொருண்மை மயக்கத்தை உருவாக்கும். எடுத்துக்காட்டாக, *கால் பகுதி.* இருப்பினும் தொடரும் வாக்கியமும் சொற்பொருண்மை மயக்கம் உள்ள சொல் இல்லாவிடினும் பொருள் மயக்கம் தரலாம். எடுத்துக்காட்டாகப் பின்வரும் தொடர்களை எடுத்துக்கொள்ளவும்.

(1) வெள்ளை மருந்துக் குப்பி

(2) பழைய மாணவர் விடுதி

முதல் வாக்கியம் ‘மருந்துள்ள வெள்ளை நிறக் குப்பி’ என்றும் ‘வெள்ளை மருந்துள்ள குப்பி’ என்றும் பொருள் மயக்கம் காட்டும். இதுபோன்று இரண்டாவது வாக்கியமும் ‘பழைய மாணவர்கள் வசிக்கும் விடுதி’ என்றும் ‘மாணவர்கள் வசிக்கும் பழைய விடுதி’ என்றும் பொருள் மயக்கம் தரும். இத்தகைய பொருண்மை மயக்கம் அமைப்புப் பொருண்மை மயக்கம் எனப்படும்.

இருவகையான பொருண்மை மயக்கங்கள் முக்கியமாகப் பேசப்படுகின்றன. அவையாவன சொற்பொருண்மை மயக்கம் மற்றும் அமைப்புப் பொருண்மை மயக்கம். இக்கட்டுரையில் சொற்பொருண்மை மயக்கம் குறித்தே ஆய இருக்கின்றோம்.

9.2.சொற்பொருண்மை மயக்கம்

சொற்பொருண்மை மயக்கம் மிகப் பொதுவானதாகும். எடுத்துக்காட்டுகள் *படி, குடி, மடம்* என்ற பெயர்களையும் *பிடி, அறை, முடி, சுடு* என்ற வினைகளையும் *நீண்ட, விரிந்த, குறைந்த, வெளுத்த, கறுத்த* என்ற பெயரடைகளையும் உள்ளடக்கும். பொருண்மை மயக்கத்திற்கு பல சோதனைகள் உள்ளன. ஒரு சோதனை *கடினமான* என்பதற்கு *மெதுவான, எளிதான* என்ற

இரண்டும் எதிர்சொற்கள் ஆகும். பின்வரும் வாக்கியத்தைக் கருத்தில் கொள்ளவும். *கடினமான மிட்டாயைக் கடிக்க முடியாது, கடினமான சொல்லுக்குப் பொருள் கூற இயலாது*. முன்னர் தந்த பொருள் மயக்கத்திற்கான எடுத்துக்காட்டு ஒரு சொல்லுக்கு ஒன்றுக்கும் மேற்பட்ட பொருண்மை காரணமாக விளைந்தது. இருப்பினும் எப்போது நமக்கு ஒரு சொல் மட்டும் இருக்கும் என்பதில் தெளிவில்லை. *படி* என்ற வினையும் *படி* என்ற பெயரும் ஒரே உச்சரிப்பைக் கொண்டிருந்தாலும் இரண்டு சொற்கள் ஆகும். இவை ஒப்புருச் சொல்லுக்கு எடுத்துக்காட்டாகும். *கடி* என்ற வினையும் *கடி* என்ற பெயரும் ஒப்புருச் சொற்களா? *முதல் மாணவனில்* வரும் *முதல்* என்பதும் *ஐந்து முதல்* என்பதில் வரும் *முதல்* என்பதும் ஒன்றா? என்ற கேள்விகள் எப்பொழுதும் எழுப்பப்படும். ஒன்று பொருள் மயக்கம் காட்டுவது மற்றொன்று ஒப்புருச் சொற்கள் என்று கூறுவதில் ஒருமித்த கருத்து இல்லை. ஒரு வேளை இந்த வேறுபாடு தன்னிச்சையானதாகும்.

சில வேளைகளில் ஒரு சொல்லின் பொருள் பிற சொல்லிலிருந்து ஆக்கப்பட்டிருக்கும். எடுத்துக்காட்டாக *காண்* என்ற வினைச் சொலின் பொருள் *கண்* என்ற சொல்லின் பொருளிலிருந்து உருவாகி இருக்கக்கூடும். இதே போன்று *எரி, உடை, ஒடி* என்ற சொற்களின் செயப்படுபொருள் குன்றாவினைப் பொருண்மைகள் செயப்படுபொருள் குன்றிய வினைப்பொருண்மைகளிலிருந்து ஆக்கப்பட்டிருக்கலாம். இப்போது இந்த நேர்வுகள் ஒவ்வொன்றிலும் ஆக்கப்பட்ட பொருண்மை அச்சொல்லின் இரண்டாவது பொருண்மையாகத் தகுதி பொறாது; ஆனால் இது ஆக்கப்படாத பொருண்மையின் மீதான சொல்லியல்சார் செயற்பாடாகும். இந்த நிகழ்வு குறிப்பிட்ட சொற்களுக்குச் சிறப்பானது என்றில்லாமல் முறையானது மற்றும் பொதுவானது என்ற அளவில் இந்த வாதம் சாத்தியமானதாகும். சொற்பொருண்மை இயலுக்கு இம்மாதிரியான பொருண்மை நிகழ்வுகளை அடையாளம் காணும் மற்றும் பண்பாக்கம் செய்யும் பொறுப்பு உள்ளது. மேலும் *செய், போடு* என்ற வினைகளும் *மேல், கீழ்* என்ற பின்னுருபுகளும் காட்டும் பொதுவான மற்றும் மிக நெகிழ்வான பொருண்மை நடத்தையை விளக்கும் பொறுப்பும் உள்ளது.

9.3.சொற்பொருண்மை மயக்கத்தின் வகைகள்

சொற்பொருண்மை மயக்கங்கள் மூன்று அடிப்படை வகைகளில் வேறு படும்: சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம் (parts-of-speech or category ambiguity), ஒப்புருச்சொன்மைசார் (சொற்போலி) பொருண்மை மயக்கம் (ambiguity due to homography), பல்பொருள் ஒருமொழியம்சார் (பல்பொருண்மை) பொருண்மை மயக்கம் (ambiguity due to polysemy), மாற்றப் பொருண்மை மயக்கம் (ambiguity due to transfer or transfer ambiguity).

9.3.1 சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம்

சொற்பொருண்மை மயக்கத்தின் நேரடியான வகை சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம் ஆகும்: ஒரு சொல் ஒன்றிற்கும் மேற்பட்ட இலக்கண அல்லது தொடரியல் வகைப்பாட்டிற்கு ஒதுக்கப்படும். தமிழில் இத்தகைய பல எடுத்துகாட்டுகள் உள்ளன. *பச்சை* என்பது பெயராகவும் பெயரடையாகவும் வரும்; *சுடு* என்பது வினையாகவும் பெயரடையாகவும் வரும்; *கடி* என்று வினையாகவும் பெயராகவும் வரும். *மேலே, கீழே* என்ற சொற்கள் பெயராகவும் வினையடைகளாகவும் பின்னுருபுகளாகவும் வரும்.

(3) அவன் மேலே இருக்கிறான். (பெயர்)

(4) அவன் மேலே சென்றான். (வினையடை)

(5) அவன் மேசை மேலே நிற்கின்றான். (பின்னுருபு)

9.3.2 ஒப்புருச்சொன்மை சார், பல்பொருள் ஒருமொழியம் சார் பொருண்மை மயக்கம்

ஒரு சொல்லுக்கு இரண்டோ அதற்கு மேலோ பொருண்மைகள் இருந்தால் இரண்டாவது வகைச் சொற்பொருண்மை மயக்கம் ஏற்படும். இரண்டோ அதற்கு மேற்பட்ட சொற்கள் முற்றிலும் வேறுபட்ட பொருண்மையைக் கொண்டிருப்பது ஒப்புருச்சொன்மை (homography) ஆகும். எடுத்துகாட்டாகப் *படி* என்பது மாடிப்படியையும் கற்பதையும் குறிப்பிடும்.

(6) அவன் தந்தை அவனிடம் நீ நன்றாகப் படி என்று கூறினார்.

(7) அவன் படி வழியாக மடியில் ஏறினான்.

அட்டை என்பது புத்தக அட்டையையும் அட்டைப் பூச்சியையும் குறிப்பிடும்.

(8) அவன் புத்தகத்தின் அட்டையைக் கிழித்து எறிந்தான்.

(9) அவன் அந்த அட்டையைக் கொன்றான்.

பல்பொருள் ஒருமொழிகள் (polysememes) தொடர்புள்ள ஒரு தொடர்ச்சியான பொருண்மைகளை வெளிப்படுத்தும். சொற்கள் தமது உருவக நீட்சியாலும் மாற்றத்தாலும் புதிய பொருண்மைகளை வெளிப்படுத்தும். எடுத்துக்காட்டாகக் *கிளை* என்பது மரக்கிளையையும் வங்கிக் கிளையையும் குறிப்பிடும்; *நட* எனபது நடக்கும் இயக்கத்தையும் நிறுவனத்தின் செயல்பாட்டையும் குறிப்பிடும்.

(10) தினமும் காலையில் பள்ளிக்கு நடந்து செல்கின்றான்.

(11) அந்த நிறுவனம் நன்றாக நடந்து கொண்டிருக்கின்றது.

ஓடு என்பது மனித இயக்கத்தையும் ஆறு ஒழுகுவதையும் குறிப்பிடும்.

(12) அவன் விரைவாக ஓடுகிறான்.

(13) காவிரி ஆறு தஞ்சாவூர் வழியாக ஓடுகின்றது.

கண் என்பது விலங்கினங்களின் கண்ணையும் தேங்காய் கண்ணையும் குறிப்பிடும்.

(14) அவன் தன் கண்களை மூடினான்.

(15) தேங்காய்க்கு மூன்று கண்கள் உண்டு.

பின்வரும் வாக்கியதில் *கேள்* என்ற சொல் செவிப்புலத்தையும் வினவுதலையும் குறித்து நின்று சொற்பொருண்மை மயக்கம் காட்டும்.

(16) ராதை ராஜா கேட்டதை அவனிடம் கூறினாள்.

இது சொற்பொருண்மைஅடிப்படையில் இரு பொருள்களைத் தந்து சொற்பொருண்மை மயக்கம் காட்டும்.

(17) ராதை ராஜா தன் காதால் கேட்டதை அவனிடம் கூறினாள்.

(18) ராதை ராஜா வினவியதை அவனிடம் கூறினாள்.

சில வேளைகளில் ஒப்புருச் சொல்லிணைகளில் ஒன்றின் பயன்பாடு மற்றதைவிட கூடுதலாக இருக்கக்கூடும். இந்நேர்வில் ஒப்புருச்சொற்களின் பொருண்மை மயக்கத்தை பனுவல் வகை அடிப்படையில் நீக்க இயலும்; எனவே வழக்கமில்லாத பயன்பாடு மொழிபெயர்க்க வேண்டிய பனுவலின் விசயத்திற்குப் பொருந்தினால் ஒழிய அகராதியிலிருந்து விலக்கி வைக்கப்படும்.

இயந்திர மொழிபெயர்ப்பில் ஒப்புருச்சொன்மையையும் (homography) பல்பொருள் ஒருமொழியத்தையும் (polysemy) பெரும்பாலும் ஒன்றுப்போல் கையாள இயலும். ஏனென்றால் இங்கு நோக்கம் எழுத்தப்பட்ட சொல்லுக்குச் சூழல் அடிப்படையில் பொருண்மையைக் கண்டுபிடிப்பது ஆகும். மேற் சொன்ன வழியில் வேறுபட்ட சொல்வகைப்பாட்டைச் சார்ந்த ஒப்புருச் சொற்களின் பொருண்மை மயக்கத்தைத் நீக்க இயலும்; ஆனால் ஒப்புருச் சொற்கள் ஒரே சொல்வகைப்பாட்டைச் சார்ந்திருந்தால் தொடரியல் பகுப்பாய்வு மட்டும் போதுமானதல்ல. பொதுவான ஒரு அணுகுமுறை 'மனிதவினம்', 'பெண்ணினம்', 'திரவம்' போன்ற பொருண்மைப் பண்புக்கூறுகளைத் தந்து தேர்வுக்கட்டுப்பாடு அடிப்படையில் எந்தப் பண்புக்கூறு தரப்பட்ட தொடரியல் கட்டுமானத்துடன் இணங்கும் என்று குறிப்பிடுவதாகும். எடுத்துக்காட்டாக குடி என்ற வினை ஒரு 'விலங்கின' எழுவாயைக் கொண்டிருக்கவேண்டும்.

தொடர்ந்து பயன்படுத்த இயலும் பொருண்மைப் பண்புகூறுகளின் குழுமத்தை உருவாக்குவதுதிலும் இப்பண்புக்கூறுகள் அடிப்படையில் பெயர்கள் மற்றும் வினைகள் இவற்றின் தேர்வுக் கட்டுப்பாடுகளைக் குறிப்பிடுவதிலும் சிக்கல்கள் உள்ளன. இருப்பினும் அவைகள் மொழிபெயர்ப்பு ஒழுங்கு முறைகளில் பெரும்பாலும் வேற்றுமைப் பங்கெடுப்பாளர்கள் (case roles) தகவலுடன் சேர்த்துப் பரவலாகப் பயன்படுத்தப்படுகின்றன. எடுத்துக்காட்டாக *அட்டை* என்ற ஒப்புருச் சொல்லை எடுத்துக்கொள்ளவும்; இதற்கு முன்னர் கூறியபடி '(புத்தக)

அட்டை' மற்றும் '(மர) அட்டை' என்ற இரு பொருண்மைகள் உண்டு. இந்த இரண்டு பொருண்மைகளையும் பின்வரும் வாக்கியங்களில் உள்ள பொருத்தமான சேர்ந்துவருகைக் கட்டுப்பாட்டுப் பண்புக்கூறுகளை (selection restriction features) விளக்கி வேறுபடுத்தலாம்.

(19) புத்தகத்தின் அட்டை கிழிந்து விட்டது.

(20) அந்த அட்டை தரையில் ஊர்ந்து கொண்டிருந்தது.

கிழி என்ற வினை *அட்டை* போன்ற கிழியக்கூடிய பொருளை எழுவாயாக ஏற்கும். *ஊர்* என்ற வினை ஊர்ந்து செல்லக்கூடிய அட்டைப் பூச்சியை எழுவாயாக ஏற்கும்.

ஒப்புருச்சொன்மை மயக்கம் இலக்கண ஆய்வு அடிப்படையிலான வேறுபாடு அடிப்பையிலும் வெளிப்படும் (இலக்கணப்போலி). பின்வரும் எடுத்துக்காட்டுகள் இதை உணர்த்தும்.

(21) அவன் கடலை தின்று மகிழ்ந்தான்

(22) அவன் கடலைக் கண்டு மகிழ்ந்தான்.

21-ஆவது வாக்கியத்தில் *கடலை* என்பது *கடலை* என்ற தின்பண்டத்தைக் குறிப்பிடும் தனிச்சொல்லாகும். 9-ஆவது வாக்கியத்தில் *கடலை* என்பது *கடல்+ஐ* எனப் பகுத்தாய இயலும் சமுத்திரத்தைக் குறிப்பிடும் திரிபுற்ற சொல்லாகும்.

பின்வரும் வாக்கியங்களில் இரு சொற்களின் வேறுபட்ட இலக்கண வடிவுகள் ஒன்றுப்போல் அமைந்து ஒப்புருச்சொன்மை காட்டுவதன் காரணமாகப் பொருண்மை மயக்கம் காட்டுகின்றன.

(23) அவன் (துணி) நெய்தான்.

(24) அவன் நெய்தான் விருப்புகிறான்

23-ஆவது வாக்கியத்திலுள்ள *நெய்தான்* என்பதும் 24-ஆம் வாக்கியத்திலுள்ள *நெய்தான்* என்பதும் இருவேறு சொற்களின் இருவேறு இலக்கண வடிவங்கள். முதல் வாக்கியம் 'he weaved (a cloth)' என்றும் இரண்டாவது வாக்கியம் 'he likes only ghee' என்றும் பொருள்பட்டு வேறுபடும்.

வரலாற்று மொழியியல் அடிப்படையில் சொற்களின் திரிபு வடிவங்கள் வேற்பட்ட இலக்கணச் செயல்பாடுகளைப் பெற்று வேறுபட்ட சொல்வகைப்பாடுகளைக் குறித்துநிற்கும் அதாவது வேறுபட்ட பொருண்மைகளைக் குறித்துநிற்கும். இருப்பினும் அவை ஒரே வடிவில் இருப்பதால் அவற்றை ஒப்புருசொற்களின் வகையாகக் கருத இயலும். குறிப்பாக வினைகளின் திரிபுற்ற வடிவங்களிலிருந்து உருவாக்கப்பட்ட அல்லது மாற்றுச் செயல்பாடு பெற்ற முன்னுருபுகள் (post postions) இத்தகைய பொருண்மை மயக்கம் காட்டும். எடுத்துக்காட்டாக

இருந்து, பற்றி, குறித்து, ஒட்டி, கொண்டு, வைத்து, சுற்றி, நோக்கி, முந்தி, விட, கூட என்ற முன்னுருபுகள் அவற்றின் மூல வினைகளின் திரிபுற்ற ஆனால் வகைப்பாடு (செயல்பாடு) மாறாத வடிவங்களுடன் பொருண்மை மயக்கம் காட்டும்.

25அ. அவன் வீட்டில் இருந்து வெளியேறினான். (முன்னுருபு)

25ஆ. அவன் வீட்டில் இருந்து வந்தான் (வினைத்திரிபு வடிவம்)

26அ. அவன் அவளைப் பற்றி பேசினான். (முன்னுருபு)

26ஆ. அவன் அவள் கையைப் பற்றி முத்தமிட்டான். (வினைத்திரிபு வடிவம்)

27அ. அவன் அவளைக் குறித்து பேசினான். (முன்னுருபு)

27ஆ. அவன் அவள் சொல்வதைக் குறித்து வந்தான். (வினைத்திரிபு வடிவம்)

28அ. அவன் அந்த தலைப்பை ஒட்டி பேசினான். (முன்னுருபு)

28ஆ. அவன் போஸ்டர் ஒட்டி பிழைக்கின்றான். (வினைத்திரிபு வடிவம்)

29அ. அவன் கத்தி கொண்டு அதை வெட்டினான். (முன்னுருபு)

29ஆ. அவன் கத்தியால் வெட்டிக் கொண்டு பேசினான். (வினைத்திரிபு வடிவம்)

30அ. அவன் கத்தி வைத்துப் பழம் வெட்டினான். (முன்னுருபு)

30ஆ. அவன் பணம் வைத்துக் கொண்டு சூதாடினான். (வினைத்திரிபு வடிவம்)

31.அ அவன் வீட்டைச் சுற்றி மரங்கள் நிற்கின்றன. (முன்னுருபு)

31ஆ. அவன் அவளையே சுற்றி வருகின்றான். (வினைத்திரிபு வடிவம்)

32அ. அவன் அவளை நோக்கி நடந்தான். (முன்னுருபு)

32ஆ. அவள் முகத்தை நோக்கிச் சிரித்தான். (வினைத்திரிபு வடிவம்)

33அ. அவன் அவளுக்கு முந்தி அங்கு வந்தான். (முன்னுருபு)

33ஆ. அவன் அவளை முந்தி நடந்துகொண்டிருந்தான். (வினைத்திரிபு வடிவம்)

34அ. அவன் அவளை விட நல்லவன். (முன்னுருபு)

34ஆ. அவன் அவளை விட விருப்பவில்லை. (வினைத்திரிபு வடிவம்)

35அ. அவன் அவள் கூட வந்தான். (முன்னுருபு)

35ஆ. அவனது மகிழ்ச்சி கூட ஆர்ப்பரித்தான்.

திரிபுற்ற *என்று* என்ற வடிவம் மூன்று வேற்பட்ட இலக்கணச் செயல்பாடுகளில் வந்து (இலக்கணப்போலி) பொருண்மை மயக்கம் காட்டும்.

36அ. அவள் நல்லவள் என்று நினைத்தேன். (என்று நிரப்பானாக வருகின்றது)

36ஆ. அவள் திடீர் என்று வந்தாள் (என்று வினையடையாக்கியாக வருகின்றது)

36இ. அவள் என்று வருகிறாள். (என்று காலத்தைக் குறிந்து வருகின்றது)

-*உம்* என்ற எதிர்கால வடிவம் வினை முற்றாகவும் பெயரெச்சாமாகவும் வரும்.

37அ. அது நாளை வரும்.

37ஆ. அது வரும் நாள் எனக்குத் தெரியாது.

-*அது* ஒட்டப்பட்ட வினைவடிவம் மூன்று வேறுபட்ட இலக்கணச் செயல்பாடுகளில் வந்து (இலக்கணப் போலி) பொருண்மை மயக்கம் காட்டும்.

38அ. அது நேற்று வந்தது (*வந்தது* வினைமுற்று வடிவம்)

38ஆ. அவன் வந்தது எனக்குத் தெரியாது (*வந்தது* வினையாலணையும் பெயர்)

38இ. அந்த செய்தித்தாள் நேற்று வந்தது (வந்தது அது மாற்றுப் பெயராக வினை)

தேர்வுக் கட்டுப்பாடு, சூழல் கட்டுப்பாடு, சூழல் அமைப்பு, சேர்ந்துவருகை என்பதன் அடிப்படையில் இத்தகைய சொற்களின் பொருண்மை மயக்கத்தை நீக்கலாம்.

9.3.3.மாற்றப் பொருண்மை மயக்கம்

வகைப்பாட்டு பொருண்மை மயக்கம், ஒப்புருச்சொன்மை, பல்பொருள் ஒருமொழியம் என்பன சொற்பொருண்மை மயக்கங்களுக்கு எடுத்துக்காட்டுகளாகும். அவை முதன்மையாக மூல மொழிப் பனுவல்களின் பகுப்பாய்வில் சிக்கலை விளைவிக்கின்றன. இயந்திர மொழிபெயர்ப்பில் மாற்றப் பொருண்மை மயக்கங்கள் (மொழிபெயர்ப்பு பொருண்மை மயக்கங்கள்) உள்ளன. ஒரு மூல மொழிச் சொல்லை இலக்கு மொழியில் பல சொற்களாலோ வெளிப்பாடுகளாலோ மொழி பெயர்க்க இயலும். மூல மொழிச் சொல் பொருண்மை மயக்கம் காட்டாது அல்லது பொருண்மை மயக்கம் இயல்மொழி பேசுபவர்களால் உணர்ந்து கொள்ளப்படாது இருக்கலாம். இச்சொல் பிற மொழி பேசுபவரின் புரிந்துகொள்ளலின் அடிப்படையில் தான் பொருண்மைமயக்கம் காட்டுகின்றது. எனவே இது மொழிபெயர்ப்பின் சிக்கலாகும்.

9.3.சொற்பொருண்மை மயக்கநீக்கம்: ஒரு கணினிசார் அணுகுமுறை

சொற்பொருண்மை மயக்கநீக்கம் என்பது ஒரு கணினிசார் வழி சூழலில் சொற்களின் பொருளைக் கண்டுகொள்ளும் ஒரு திறனாகும். சொற்பொருண்மை மயக்கநீக்கம் ஒரு செயற்கை அறிவு-முழுமைச் சிக்கலாகக் (AI-complete problem) கருதப்படுகின்றது; அதாவது, செயற்கை அறிவில் மிகக் கடினமான சிக்கல்களுக்குச் சமமான செயல்பாடாகும். இங்கே சொற்களின் மயக்கத்தை நீக்குவதற்கான செயலூக்கங்களும் இச்செயலின் விளக்கமும் தரப்படும். இங்கு கண்காணிக்கப்பட்ட, கண்காணிக்கப்படாத மற்றும் அறிவு அடிப்படையிலான அணுகுமுறைகள் சுருக்கமாகக் கூறப்படும்.

9.3.1 புற அறிவு மூலங்கள்

சொற்பொருண்மை மயக்கநீக்கத்தின் அடிப்படைக் கூறு/பகுதி அறிவு ஆகும். அறிவு மூலங்கள் சொற்களுடன் பொருண்மைகளைத் தொடர்புபடுத்தத் தேவையான தரவுகளைத் தருகின்றன. அவை புலக்குறிப்பு செய்யப்படாத (unlabeled) அல்லது அடையாளப்படுத்தப்பட்ட (annotated) பனுவல்களின் தரவுத்தொகுதிகளிலிருந்து (corpora of texts) இயந்திரத்தால் படிக்கவியலும் அகராதிகள் (machine readable dictionaries (MRDs)), பொருட்புல அகராதிகள் (thesauri), பொருள்விளக்கச் சொற்கோவை (glossaries), மூலப்பொருண்மையியல் ஆய்வுகள்/மெய்ப்பொருள் மூலாய்வு (ontologies) மற்றும் பிற என வேறுபடலாம்.

அமைப்பாக்கம் செய்யப்பட்ட மூலவளங்கள் (structured resources):

- பொருட்புல அகராதிகள்
- இயந்திரத்தால் படிக்கவியலும் அகராதிகள்
- மூலப்பொருண்மையியல் ஆய்வுகள் (Ontologies):

அமைப்பாக்கம் செய்யப்படாத மூலவளங்கள் (unstructured resources):

- தரவுத்தொகுதிகள் (corpora):
- கச்சாத் தரவுத்தொகுதிகள் (raw corpora)
- பொருண்மை அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் (sense-annotated corpus)
- சேர்ந்துவருகை மூலவளம் (collocational resources)
- பிற மூலவளங்கள்
- சொல்வலை
    - பொருள்விளக்கம் (gloss)
    - சொல்லன்சார் மற்றும் பொருண்மைசார் உறவுகள் (Lexical and semantic relations)
        - எதிர்மொழியம் (antonymy):
        - சார்பு மொழியம் (Pertainymy):
        - பெயராக்கம் (Nominalization):
        - உள்ளடக்கு மொழியம் (hypernymy)
        - உள்ளடங்கு மொழியம் மற்றும் உட்படு மொழியம் (Hyponymy and troponymy):

9.3.2 கணினிசார் பொருண்மை மயக்க நீக்கத்தின் வகைகள்

கணினிசார் பொருண்மை மயக்கநீக்கதை மூன்றாகப் பிரிக்கலாம்.

1.கண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம்

2. கண்காணிக்கப்படாத பொருண்மை மயக்கநீக்கம்

3. அறிவு அடிப்படையிலான பொருண்மை மயக்கநீக்கம்

9.3.2.1. கண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம்

கண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம் பின்வருவனவற்றை உள்ளடக்கும்.

1. தீர்மான பட்டியல்கள்
2. தீர்மானக் கிளைகள்
3. நெய்வ் பெய்ஸ்
4. நரம்புசார் வலையமைப்பு
5. முன்மாதிரி அடிப்படையிலான அல்லது எடுத்துக்காட்டு அடிப்படையிலான கற்றல்
6. சப்போர்ட் வெக்டார் மெஷியன்கல்
7. ஒருங்கிணைந்த நெறிமுறைகள்
    - பெரும்பான்மை வாக்களிப்பு
    - நிகழ்தகைமைக் கலப்பு
    - தர அடிப்படையிலான ஒருங்கிணைப்பு
    - அடாபூஸ்ட்

8.மிகக்குறைவான மற்றும் பகுதி கண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம்

9.3.2.2 கண்காணிக்கப்படாத பொருண்மை மயக்கநீக்கம்

கண்காணிக்கப்படாத பொருண்மை மயக்கநீக்கம் பின்வருவனவற்றை உட்படுத்தும்.

1. சூழல் கொத்தாக்கம்
2. சொல் கொத்தாக்கம்
3. சேர்ந்துவருகை வரைபடங்கள்

9.3.2.3 அறிவு அடிப்படையிலான பொருன்மை மயக்கநீக்கம்

அறிவு அடிப்படையிலான பொருண்மை மயக்கநீக்கம் பின்வருவனவற்றை உட்படுத்தும்.

1. பொருண்மை மயக்கநீக்கங்களின் மேலுறல்
2. விருப்பத்தேர்வுகள்
3. அமைப்புமுறை அணுகுமுறைகள்
    - ஒற்றுமை அளவீடுகள்
    - வரைபட அடிப்படையிலான அணுகுமுறைகள்

9.3.2.4 பிற அணுகுமுறைகள்

9.4. தமிழில் சொற்பொருள் மயக்கநீக்கத்திற்கான நெறிமுறைகள்

ஒரு பனுவலில் உள்ள சொல் மயக்கங்களை நீக்க ஒரு பொருண்மை ஆய்வி தேவையாகும். இந்த ஒழுங்குமுறைத் தமிழை மூலமொழியாகக்கொண்ட இயந்திர மொழிபெயர்ப்புக்கும் தமிழ் தரவுகளின் அறிவாற்றல் உள்ள தகவல் மீட்பு ஒழுங்குமுறைக்கும் தேவையானதாகும். எடுத்துக்காட்டாகத் தமிழ்-ஆங்கில இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை பின்வரும் வாக்கியத்தை உள்ளீடாகக் கொள்கின்றது என்று கொள்வோம்.

39. நீலகிரி காடுகளில் அரிய வகை விலங்குகள் காணப்படுகின்றன.

மேற்கண்ட வாக்கியத்தில் *விலங்குகள்* என்பது *விலங்கு* என்பதன் பன்மை வடிவமாகும். *விலங்கு* என்பதற்கு இரண்டு வேறுபட்ட அர்த்தங்கள் உள்ளன:

1. மிருகம்
2. கையில் போடும் விலங்கு

'விலங்கு' அர்த்தம் '(கை)விலங்கு' அர்த்தத்தைவிட அதிக நிகழ்வெண்ணில் வருகின்றது. ஆங்கிலத்தில் இந்த இரு அர்த்தங்களுக்கும் தனித்தனியான வடிவச் சொற்கள் உள்ளன: விலங்கு1 'animal', விலங்கு2 'hand-cufs'. எனவே பொருண்மையியல் ஆய்வி இப்பொருண்மை மயக்கத்தை நீக்க வேண்டும்; இதன் படி மேற்சொன்ன தமிழ் வாக்கியம் பின்வருமாறு மொழிபெயர்க்கப்பட வேண்டும்.

40. Rare animal species are found in Nilgiris forests.

கீழே எடுத்துக்காட்டாக இரு அணுகுமுறைகள் விவரிக்கப்பட்டுள்ளன.

9.4.1 கொத்தாக்கச் சேர்ந்துவருகை அணுகுமுறை

இவ்வாய்வில் பொருண்மை மயக்க நீக்கம் கொத்தாக்க உபாயம் (clustering technique) மற்றும் அறிவு மூலவளங்களாகச் சேர்ந்துவருகைகளையும் (collocations) வேற்றுமை குறியீடுகளையும் பயன்படுத்துவது அடிப்படையில் நடைமுறை படுத்தப்பட்டுள்ளது. ஒழுங்குமுறையின் செயல்பாடு படத்தில் தரப்பட்டுள்ள ஒழுங்குமுறை ஒழுக்குப் படத்தில் காட்டப்பட்டுள்ளது.

இவ்வணுகுமுறை இரண்டு கட்டங்களில் செயல்படுகின்றது. பயிற்சி கட்டத்தின் தொடக்கத்தில் ஒழுங்குமுறை பொருண்மை மயக்கம் உள்ள சொற்களைக் கண்டுபிடிக்கின்றது. வேறுபட்ட ஆய்வாளர்கள் வேறுபட்ட அளவுகளில் சூழல்களைப் (அதாவது வேறுபட்ட சாளர அளவுகளில் (different window size)) பயன்படுத்தியுள்ளனர். சாளர அளவு ஒழுங்குமுறையின் செயல்பாட்டைக் கணிசமான அளவில் பாதிக்கின்றது. ஒரு பெரிய சாளரம் சூழலுக்கு அதிக

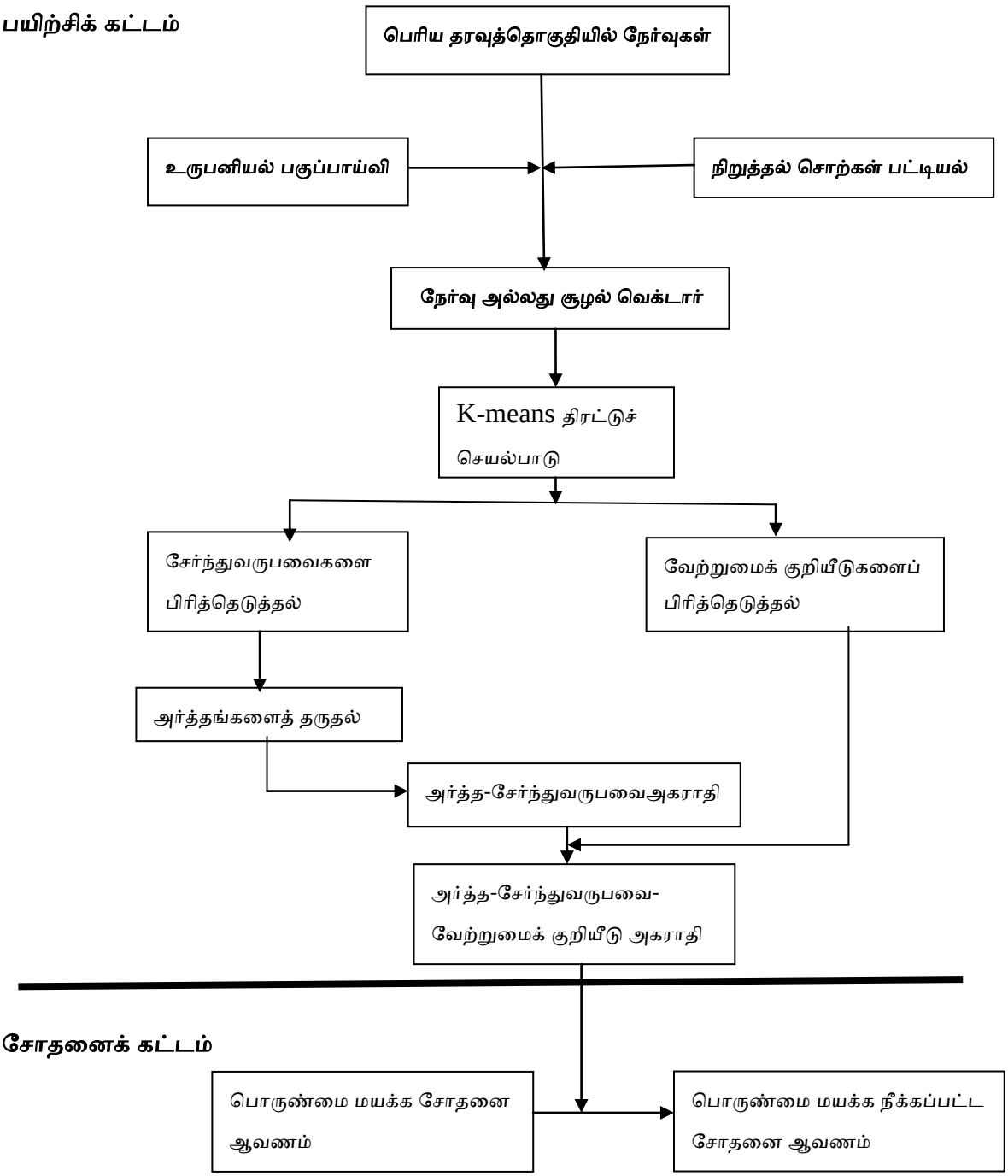
அளவில் தொடர்பில்லாத சொற்களைக் கொண்டுவருகின்றது; ஆனால் சிறிய சாளரம் சில முக்கியமான சேர்ந்துவருபவைகளை இழக்கின்றது. யரொவ்ஸ்கி (Yarowsky 1995) உட்பட பல சொற்பொருண்மை மயக்கநீக்கிகளில் காணப்படும் பொதுவான நடமுறையைப் பின்பற்றி இவ்வாய்வில் இருபது-சொல் சாளரம் பயன்படுத்தப்பட்டுள்ளது.

நிறுத்தல்-சொற்கள் (stop-words) என்று அழைக்கப்படுகின்ற அதிக நிகழ்வெண் உள்ள சொற்கள் சாளரத்திலிருந்து நீக்கப்படும். பின்னர் திரட்டுவதில் பயன்படுத்தப்படும் சூழல் சொற்களின் வேர்வடிவைப் பெற உருபனியல் அடிப்படையில் ஆயப்படும். ஒவ்வொரு சூழலும் சூழல் அல்லது நிகழ்வு வெக்டராகப் உருப்படுத்தம் செய்யப்படும். சேர்ந்துவருபவைகள் தானியக்கமாக இந்தத் திரட்டுகளிலிருந்து சேகரிக்கப்படும் மற்றும் அர்த்தங்கள் மனித-அடையாளப்படுத்திகளால் (human-annotators) ஒதுக்கப்படும்/தரப்படும்; இதன்படி ஒரு அர்த்த-அகராதி உருவாக்கப்படும். வேற்றுமை-குறியீடுகள் ஒவ்வொரு திரட்டுக்கும் அடையாளம் காணப்படும் மற்றும் அர்த்த-சேர்ந்துவருபவை வேற்றுமை-குறியீடு அகராதியை (sense-collocation case-marker dictionary) உருவாக்க வேண்டி சேர்க்கப்படும்.

இரண்டாவது கட்டத்தில் ஒழுங்குமுறை பொருண்மை மயக்கநீக்கத்திற்குத் தயாராக இருக்கும். பொருண்மை மயக்கச் சொற்கள் கொண்ட எந்தப் பனுவலும் ஒழுங்குமுறைக்குத் தரப்படலாம். படத்தில் காட்டியுள்ளது போன்று ஒழுங்குமுறை பொருண்மை மயக்கச் சொற்களின் பார்க்கப்படாத நிகழ்வுகளை பொருண்மை மயக்கநீக்கம் செய்ய அர்த்த-சேர்ந்துவருகை வேற்றுமை-குறியீட்டு அகராதியைப் (Sense-Collocation-Case marker dictionary) பயன்படுத்துகின்றது. ஒவ்வொரு பொருண்மை மயக்கம் உள்ள சொற்களுக்கும் பயிற்சி ஒருநேரச் செயல்பாடாகும். ஒழுங்குமுறை ஒரு குறிப்பிட்ட சொல்லுக்காகப் பயிற்சி பெற்றுவிட்டால் அதை ஒரு தரப்பட்ட சூழலில் அச்சொல்லின் பொருண்மை மயக்கத்தை நீக்கப் பயன்படுத்த இயலும். இந்த நெறிமுறையில் சேர்துவருபவைகளும் வேற்றுமைக் குறியீடுகளும் ஒரு முக்கியமான பங்களிப்பு செய்கின்றது என்பதைக் காணலாம்.

ஒழுங்குமுறை ஒழுக்குப் படம்

பயிற்சிக் கட்டம்



சோதனைக் கட்டம்

9.4.1.1 சேர்ந்துவருபவைகளின் முக்கியத்துவம்

சேர்ந்துவருபவைகள் தந்துள்ள ஒரு வாக்கியத்தில் பொருண்மை மயக்கச் சொற்களின் அர்த்தத்தை உறுதியாக அனுமானிக்கும் அடுத்துவரும் சொற்கள் ஆகும். பொதுவாகச் சேர்ந்துவருபவை இரண்டு சொற்களுக்கு இடையில் அளக்க இயலும் இட-சிறப்பு உறவுகளைக் குறிப்பிடுகின்றது. சேர்ந்துவருபவைகள் பொருண்மை மயக்கமுள்ள சொற்களுக்குப் பொருண்மையியல் அடிப்படையில் மிக அண்மையில் வரும் சொற்களைப் பற்றிய குறியாக்கம் செய்யப்பட்ட தகவலாகும். எடுத்துக்காட்டாகப் பின்வருமாறு இரு அர்த்தங்கள் உள்ள plant என்ற சொல்லைக் கருத்தில் கொள்ளவும்.

plant1 – a living organism

plant2 – a manufacturing place

plant என்ற சொல்லின் சேர்ந்துவருபவைகள் பனுவலில் *plant* என்பதுடன் அடிக்கடிச் சேர்ந்துவரும் சொற்களை உட்படுத்தும். *plant* என்பதன் மேற்கண்ட அர்த்தங்களுக்கு எடுத்துக்காட்டான சேர்ந்துவருபவைகள் கீழே பட்டியலிடப்பட்டுள்ளன.

plant1 – growth, height, flower, fruit, species, leaves

plant2 – car, union, equipment, assembly, nuclear, job, worker

பெரும்பாலும் சேர்ந்துவருபவைகள் பொருண்மை மயக்கமுள்ள சொற்களிருந்து உள்ள தூரத்திற்கு உணர்வுள்ளவைகளாகும். பொருண்மை மயக்கச் சொற்களிலிருந்து குறைந்த தூரத்தில் உள்ள சொற்கள் வலுவாக அதன் அர்த்தத்தைச் சுட்டிக்காட்டும். தூரம் அதிகரிக்கும் போது பொருண்மைமயக்கச் சொற்களுடன் உள்ள அதன் சம்பந்தம் குறையும். தூரம் பொதுவாகச் சாளர அளவு என்று குறிப்பிடப்படும். இருப்பினும் சேர்ந்துவருவது என்று கருதுவதற்கு ஒரு சொல் பொருண்மைமயக்கம் உள்ள சொல்லிலிருந்து இருக்கவேண்டிய அறுதியான தூரம் இல்லை. சில வேளைகளில் சேந்துவருபவைகள் பொருண்மைமயக்கச் சொற்களிலிருந்து மிகுந்த தூரத்தில் வரலாம். தற்போதைய அணுகு முறையில் சேர்ந்துவரும் சொற்கள் யாவும் தானியக்கமாகத் திரட்டப்பட்ட நேர்வுகளிலிருந்து பிரித்தெடுக்கப்படும். இது விதை சேர்ந்துவருபவைகளைக் (seed collocations) கையால் தரும் தேவையை நீக்குகின்றது; ஒரு மொழியில் வரும் பொருண்மை மயக்கமுள்ள சொற்களைக் கருத்தில் கொள்ளும்போது இது சிரமமான வேலையாகும்.

9.4.1.2 வேற்றுமை குறியீடுகளின் முக்கியத்துவம்

ஆங்கிலத்தில் பெயருக்கும் வினைக்கும் உள்ள உறவுகள் *in, on. at, by, with* போன்ற முன்னுருபுகளால் வெளிப்படுத்தப்படும். தமிழில் இது பின்னுருபுகளால் வெளிப்படுத்தப்படும்.

வர்ணம் என்ற தமிழ்ச் சொல்லை எடுத்துக்கொள்வோம். இதற்கு குறைந்தது மூன்று வேறுபட்ட அர்த்தங்கள் உள்ளன.

1. நிறம்
2. இந்திய மரபு இசையில் ஒரு குறிப்பிட்ட பாடல் வகை
3. பண்டை இந்தியாவில் பிறப்பு அடிப்படையிலான பகுப்பு

இவ்வர்த்தங்கள் வாக்கியங்களில் வருகையில் வேற்றுமை உருபுகளின் தேர்வு சில வேளைகளில் இப்பொருண்மை வேறுபாடுகளை வெளிப்படுத்தலாம். *மாலை* என்ற சொல்லை எடுத்துக்கொள்வோம். இது மாலைக் காலத்தையும் கழுத்தில் அணியும் மாலையையும் குறிப்பிட்டுப் பொருண்மை மயக்கம் காட்டும். இவை பின்வருமாறு வாக்கியங்களில் பயன்படுத்தப்படலாம்.

மாலையில் பள்ளி ஆண்டுவிழா கொண்டாடப்பட்டது.

அவன் கழுத்தில் மாலை அணிவித்தார்கள்.

மேற்கண்ட வாக்கிய நேர்வுகளில் வேற்றுமை உருபுகள் அர்த்தங்களை வேறுபடுத்த உதவக்கூடும்.

9.4.1.3 கற்றல் கட்டம்

கொத்தாக்கல்/திரட்டல் செயற்பாங்கு (clustering process) இங்கு விளக்கப்பட்டுள்ளது.

சூழல்களும் சூழல் இடைவெளியும்

ஒரு ஆவணத்தில் சொற்பொருள் மயக்கமுள்ள சொல்லின் ஒவ்வொரு நேர்வும் அதன் சூழல் என்று அழைக்கப்படும். ஒரு பெரிய தரவுத்தொகுதியில் சொற்பொருண்மை மயக்கம் உள்ள சொல்லின் எல்லா சூழலகளும் S என்ற கருத்துச்சாயலால் குறிப்பிடப்பட்ட சூழல் இடைவெளியில் (context space) உருப்படுத்தம் செய்யப்படும். எடுத்துக்காட்டாக *மாலை* என்ற சொல்லைக் கருத்தில் கொள்வோம். 'மாலை (காலம்)' என்ற அர்த்தம் S1 என்றும் '(கழுத்திலிடும்) மாலை' என்ற அர்த்தம் S2 என்றும் கொள்வோம். மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்ட மூன்று மில்லியன் தமிழ்த் தரவுத்தொகுதியிலிருந்து பிரித்து எடுக்கப்பட்ட வாக்கியங்களில் வரும் *மாலை* என்பதை இரு அர்த்தங்களும் அடையாளப்படுத்தப்பட்டு கீழே தரப்பட்டுள்ளன.

தெரசாவின் அறப்பணிகளுக்கு நிகரில்லாத புகழ் மாலைகளாக அமைந்து மகிழ்வித்தன. S2

…தினந்தோறும் காலை மாலை ஆகிய இரு வேளைகளிலும் S1

காலையில் சூரியம் தோன்றிய நேரத்தில் இருந்து மாலையில் மறையும் நேரம் வரை S1

சூரியகாந்திப்பூ காலை, மாலை வேளைகளில் திசை S1

சுற்றுலாப் பேருந்து அடுத்த வெள்ளிக்கிழமை மாலை சென்னைக்குத் திரும்புகிறது S1

காலையில் 5 மணிக்கும், பிறகு 9:30 மணிக்கும், மலையில் 6:30 மணிக்கும் S1

அரசு தூக்கிலிட முடிவு செய்த நாளின் முதல் நாள் மலையே, தூக்கிலிட்டுக் S1

தத்தும் வீசிய துண்டறிக்கைகள் அன்று மாலையே "இந்துஸ்தான் டைம்ஸ்" S1

நடவடிக்கைகளின் காரணமாக நேற்று மாலையிலேயே வந்த செய்தியின்படி S1

"குரங்கு கையிலே பூமாலை" என்று அதைப்போல ஊராட்சி S2

சிலர், 'நான் மலையையே மலர் ஆக்குவேன் S2

அழகு, இளமை, கருவிழி, செல்லம், வாகனம், மாலை, கொடி முதலியவை S2

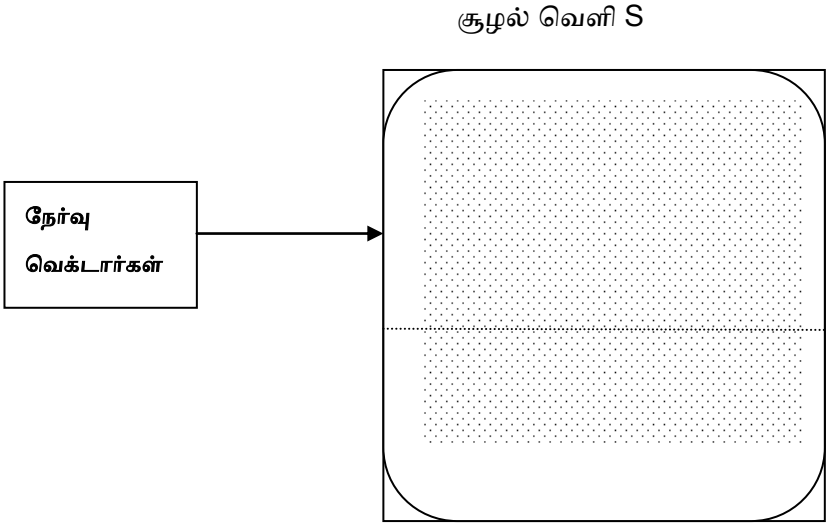
அல்லது திருநீறு காணப்படும் சபையோர் போடும் மாலைகளைப் கழற்றாமலே S2

நானும் அந்தச் செய்தியை நேற்று மாலை அமைச்சரவைக் கூட்டம் S1

உணவுப் பொருள்களை உட்கொள்ளும் பொழுதும், மாலையில் மேய்ச்சலுக்குப் பிறகும் S1

'மாலை (காலம்)' S1என்ற அர்த்தத்தில் வந்த எல்லா சூழல்களும் படத்தின் மேல்பாகத்திலும் '(கழுத்தில் அணியும்) மாலை' S2 என்ற அர்த்தத்தில் வந்த எல்லா சூழல்களும் படத்தின் கீழ்ப் பகுதியிலும் காட்டப்பட்டுள்ளன. பிரிக்கக்கூடிய புள்ளிக் கோடு இவ்விரு அர்த்தங்களின் சூழல் வெளியைக் காட்டுவதற்காக இடப்பட்டுள்ளது. கொத்தாக்கல்/திரட்டுச் செயற்பாங்குகள் வேறுபட்ட சுழல்களின் உண்மையான பிரிவை எய்துகின்றது.

சொற்பொருண்மை மயக்கம் உள்ள சொல்லின் ஒவ்வொரு சூழலும் சூழல் வெக்டார்களாகக் கையாள இயலும்; அவை சூழல் வெளியில் உருப்படுத்தம் செய்யப்படும். சூழல் வெக்டாரின் தனிமங்கள் பொருள்மயக்கம் உள்ள சொல்லின் எல்லா சூழல்களிலும் தோன்றும் எல்லா சொற்களையும் உட்படுத்தும். சூழல்களை சூழல் வெக்டார்களாக உருப்படுத்தம் செய்யும் நெறிமுறை ஒரு எடுத்துக்காட்டால் கீழே காட்டப்பட்டுள்ளது.



இப்போது பொருள்டக்கச் சொற்களை அடையாளம் காண்பதன் செயற்பாங்கை விளக்க மூன்று சூழல்களைக் கருத்தில் கொள்ளவும்.

C1 – மாலையில் நடைபெற்ற விழாவில் மாணவர்களுக்குப் பரிசு வழங்கப்பட்டது

C2 – வரவேற்பு விழா மாலை சரியாக ஆறு மணிக்குத் துவங்கியது.

C3 – ரோஜாப்பூ மாலை இருபது ரூபாயில் கிடைக்கிறது

சூழல் சாளரத்தில் வரும் எல்லாச் சொற்களும் உருபனியல் அடிப்படையில் ஆயப்பட்டது மற்றும் விளையும் வேர்வடிவுகள் திறனுள்ள பொருளட்டக்கச் சொற்களாகக் கருதப்படும். இவை நிறுத்தல் சொற்களை நீக்கி மேலும் குறைக்கப்படும். C1-க்கு ஆறு பொருளடக்கச் சொற்களைக் கண்டுபிடிக்க இயலும்.

மாலை, நடைபெற்ற, விழா, மாணவர், பரிசு, வழங்கு

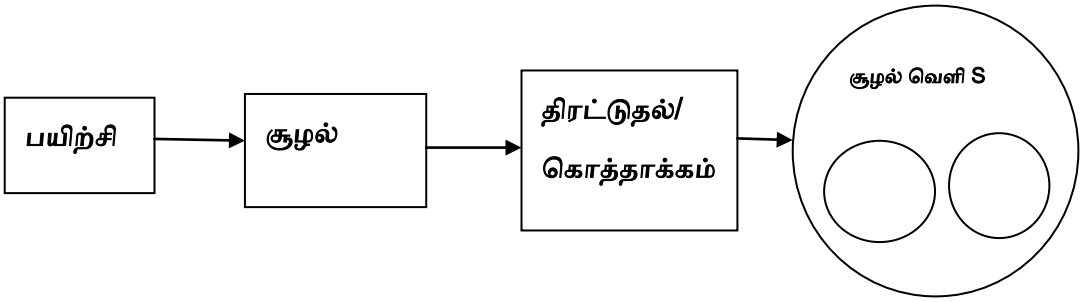
இதே வழியில் தரவுத்தொகுதியில் வரும் எல்லா நேர்வுகளுக்கும் பொருளடக்கச் சொற்களைக் கண்டுபிடிக்க இயலும். இவ்வாறு நமது குறைந்த மூன்று நேர்வுகளுக்கு 5 பொருளடக்கச் சொற்களைக் கண்டுபிடிக்க இயலும்.

மலை, நடைபெற்ற, விழா, மாணவர், பரிசு, வழங்கு, வரவேற்பு, ஆறு, மணி, துவங்கு, ரோஜாப்பூ, இருபது, ருபாய், கிடைக்கிறது

ஒவ்வொரும் நேர்வும் ஒரு சூழல் வெக்ட்டராக இப்போது உருப்படுத்தம் செய்யப்படும்; இதைச் சூழல் வெளியில் ஒரு தனிப்பட்ட இடத்தில் காட்ட இயலும். சூழல் வெக்டாரின் நீளம் தரவுத்தொகுதியில் எல்லா நேர்வுகளுக்கும் அடையாளம் கண்டுபிடிக்கக்கூடிய பொருளடக்கச் சொற்களின் எண்ணிக்கை ஆகும். இவ்வாறு எடுத்துக்காட்டாக *C1*, *C2* மற்றும் *C3* முன்று

சூழல்களுக்கு ஒவ்வொரு சூழல் வெக்டாரின் நீளம் 14 ஆகும். சூழல் வெக்டாரில் உள்ள ஒவ்வொரு தனிமமும் அந்த குறிப்பிட்ட சொல் அச்சூழலில் இருக்கிறதா அல்லது இல்லையா என்பதைப் பொறுத்து பூஜியத்தையோஅல்லது ஒன்றையோ எடுக்கும். திரட்டும் வழிமுறை வரைவு பொருண்மை மயக்கம் உள்ள சொல்லின் நேர்வுகளை வேறுபட்ட திரட்டுகளாகக் குழுமும்; ஒவ்வொரு கொத்தும்/திரட்டும் சூழல் வெளியில் ஒரு குறிப்பிட்ட இடத்தில் செறிவாக இருக்கும். கொத்தாக்கத்தின்/திரட்டுவதன் அடிப்படைக் கருத்து ஒரு கொத்தில்/திரட்டில் சூழல் அடிப்படையில் ஒற்றுமையுள்ள நேர்வுகளைக் குழுமுவதாகும். பின்வரும் படம் இவ்வணுகுமுறையில் கொத்தாக்கத்தின்/திரட்டும் செயற்பாங்கின் செயல்பாட்டைத் தெளிவாகக் காட்டுகின்றது. இது பொருண்மை மயக்கநீக்கதிற்கு வேண்டி இங்குப் பயன்படுத்தப்பட்டுள்ள திரட்டும் உபாயத்தை எடுத்துக்காட்டுகின்றது.

தரவுத்தொகுதியிலிருந்து பொருண்மை மயக்கச்சொற்களின் எல்லாச் சூழல்களும் முதலில் சூழல் வெக்டார்களாக உருப்படுத்தம் செய்யப்படும். பின்னர் சூழல் வெக்டார்கள் எல்லாம் K-மீன்ஸ் வழிமுறைவரைவைப் பயன்படுத்தித் திரட்டப்படும்.



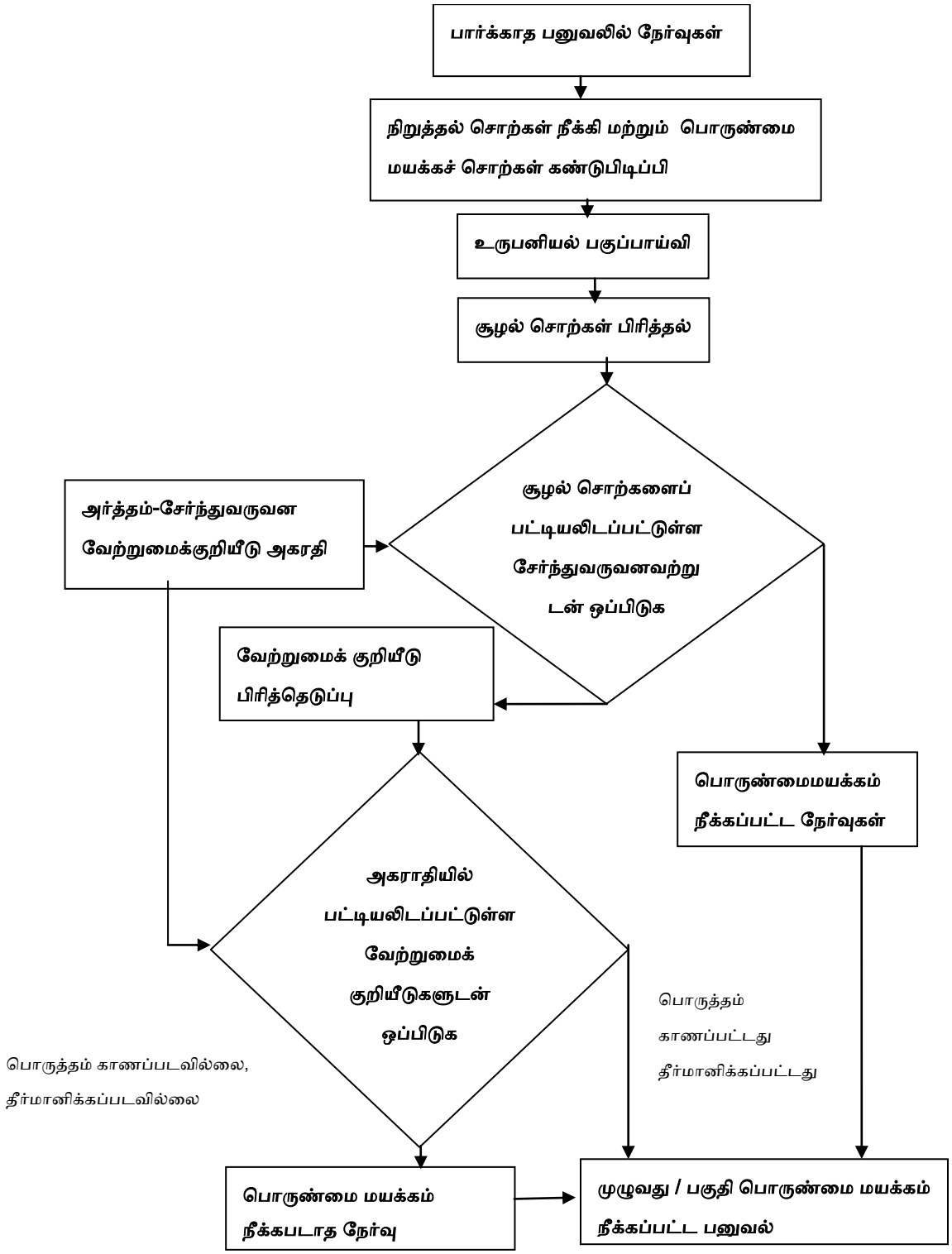
கொத்தின்/திரட்டின் இறுதியில் சூழல் அடிப்படையில் ஒத்திருக்கின்ற கொத்துக்/திரட்டுக் குழுக்கள் பெறப்படும். கொத்துக்களிலிருந்து/திரட்டுகளிலிருந்து சேர்ந்துவருபவைகளைப் பிரித்தெடுத்தபின் அடுத்த நடவடிக்கை அர்த்தம் அடையாளப்படுத்துவது ஆகும். இங்கு மனித அடையாளப்படுத்துபவர்கள் உயர்ந்த பத்து சேர்ந்துவருகைகளும் பொருண்மை மயக்கச் சொற்களின் அர்த்தங்களும் தரப்படுவார்கள். சொற்கள் வேறுபட்ட குழுமங்களாகப் பிரிக்கப்படும் மற்றும் ஒவ்வொரு குழுமமும் வேறுபட்ட அடையாளப்படுத்துபவரிடம் கொடுக்கப்படும். ஒவ்வொரு சேர்ந்துவரும் குழுமத்திற்கும் கூடுதல் பொருத்தமான அர்த்தத்தைத் தொடர்புபடுத்தி ஒரு அர்த்தம்-சேர்ந்துவருகை அகராதி உருவாக்குவதாகுவது அடையாளப்படுத்துபவர்களின் வேலையாகும்.

சொல்	அர்த்தம்	சேர்த்துவரும் சொல்
மாலை	மாலைப் பொழுது	காலை, பகல், சூரியன், மறை, நேரம், நாள், மாதம், மதியம், இரவு
மாலை	(கழுத்திலிடும்) மாலை	பூ, மணம், ரோஜா, மல்லிகை, மலர், பூஜை, மகளிர், கொன்றை
நூல்	நூல் (இழை)	இழை, தறி, சாயம், நெய், கதர், நெசவு, ஊடு, பாவு, துணி, ஊசி
நூல்	புத்தகம்	ஆசிரியர், பதிப்பகம், அச்சு, பக்கம், தாள்
விலங்கு	மிருகம்	காடு, சிங்கம், புலி, தோல், யானை, தந்தம், நாய், பூனை
விலங்கு	(கை) விலங்கு	குற்றவாளி, கோர்ட், போலீஸ், கைதி, விசாரணை, தண்டனை

9.4.1.4 பரிசோதனைக் கட்டம்

இப்போது ஒழுங்குமுறை பயன்பத்துபவர் பயன்படுத்துவதற்குத் தயாராக உள்ளது. இதில் ஒருவர் பொருண்மை மயக்கச் சொற்கள் உள்ள தமிழ்ப் பனுவலை உள்ளீடாகக் கொடுக்க இயலும் மற்றும் முழுவதும் அல்லது பகுதி பொருண்மை மயக்கம் நீக்கப்பட்ட பனுவலை வெளியீடாகப் பெற இயலும். பரிசோதனைக் கட்டத்தில் அர்த்தம்-சேர்ந்துவருகை வேற்றுமை குறியீட்டு அகராதி பார்க்கப்படாத நேர்வுகளைப் பொருண்மை மயக்கம் நீக்குவதற்குப் பயன்படுத்தப்படுகின்றது. பரிசோதனைக் கட்டத்தில் கண்டுபிடிக்கப்படும் சேர்ந்துவருவனவும் வேற்றுமைக் குறியீடுகளும் தரப்பட்ட பனுவலில் உள்ள பொருண்மை மயக்கங்களை நீக்கப் பயன்படுத்தப்படுகின்றது. இருப்பினும் வழிமுறை வரைவு தரப்பட்ட பனுவலில் தோன்றும் பொருண்மை மயக்கமுள்ள சொற்களுக்கும் முன்னரே பயிற்சிசெய்யப்பட்டிருக்கவேண்டும்.

Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019



9.4.2. கருத்துரு வரைபட அகராதி அடிப்படையில் சொற்பொருண்மை மயக்க நீக்கம்

இவ்வாய்வில் சொற்பொருண்மை மயக்கநீக்கத்திற்கு வேண்டி கருத்துரு வரைபட அகராதி/ கருத்துரு அகராதி உருவாக்கப்பட்டுப் பயன்படுத்தப்பட்டுகின்றது. கருத்துரு வரைபட அகராதியின் உருவக்கம் பற்றியும் அதைப் பயன்படுத்திச் சொற்பொருண்மை மயக்க நீக்கம் எவ்வாறு மேற்கொள்ளப்படுகின்றது என்றும் இங்கு விரிவாக விளக்கபட்டுள்ளது. இதற்கு அடிப்படையாக அமையும் சொற்பொருண்மை ஆய்வு (lexical semantics study) பற்றி முன்னர் விளக்கப்பட்டுள்ளது. இது அறிவு அடிப்படையிலான நெறிமுறையை/அணுகுமுறையைச் சாரும். இந்நெறிமுறையில் கருத்துரு வரைபட அகராதி சொற்களின் பொருண்மை மயக்கநீக்கத்திற்குப் பயன்படுத்தப்படுகின்றது. இது ஒரு மாதிரி (model) நிலையிலேயே அமைந்துள்ளது. பொருண்மை மயக்கம் தருகின்ற சில சொற்கள் தேர்ந்தெடுக்கப்பட்டுத் தரவுத்தொகுதியிலிருந்து அச்சொற்கள் வரும் வாக்கியங்கள் பிரித்தெடுக்கப்பட்டு பொருண்மை மயக்கநீக்கம் செய்ய முயற்சி எடுக்கப்பட்டுள்ளது. இதற்கு என்று அமிர்தாப் பல்கலைக்கழகத்தின் தரவுத்தொகுதி பயன்படுத்தப்பட்டுள்ளது முழுச் சொற்றொகைக்கும் தற்போது இதைப் பயன்படுத்த இயலாது. இருப்பினும் ஒரு முழுமையான கருத்துரு வரைபட அகராதி உருவாக்கப்பட்டால் இது ஓரளவுக்குச் சாத்தியமாகும்.

இது இருகட்டங்களைக் கொண்டது: (1) கற்றல் கட்டம் (2) பரிசோதனைக் கட்டம். முதல் கட்டத்தில் கருத்துரு வரைபட அகராதியைப் பயன்படுத்தி தரவுத்தொகுதியில் வரும் பொருண்மை மயக்கமுள்ள சொற்கள் வேறுபட்ட கருத்துருக்களுக்காக (அதாவது பொருண்மைகளுக்காக/அர்த்தங்களுக்காக) அடையாளப் படுத்தப்படும். இரண்டாம் கட்டத்தில் அடையாளப்படுத்தப்படாத தரவுத்தொகுதியில் உள்ள செற்களின் பொருண்மை மயக்கங்களை நீக்கப் பரிசோதனை செய்யபடும்.

9.4.2.1 கருத்துரு வரைபடம்

செயற்கை அறிவு வெளிப்படுத்தத்திற்குப் பயன்படுத்தப்படும் கருத்துரு சார்ந்த அமைப்பு அடிப்படையிலான கோட்பாடு சௌவா (Sowa, 1984) என்பவரால் விளக்கப்பட்டுள்ளது. இது மூளையிலும் இயந்திரத்திலும் (கணிப்பொறியிலும்) தகவலை ஆய்வதற்கான சாத்தியம் அடிப்படையில் அமைந்தது. கருத்துரு வரைபடம் நபர்களுக்கும் (persons) பொருளுக்கும் (things) நிகழ்வுகளுக்கும் (events) இடையிலுள்ள உறவுகளைக் கூறும் தர்க்க வடிவங்களாகும். சாம்ஸ்கி மாற்றிலக்கணம் அறிதிறனின் ஒரு அருவக் கோட்பாடு என்றும் அது செயல் திறனின் கோட்பாடாகக் கருதப்படக்கூடாது என்பார். ஆனால் செயற்கை அறிதிறனுக்கு

வழியமைப்பாளர்கள் மக்களுக்கும் கணிப்பொறிக்கும் இடையில் நடக்கும் கருத்துப்பரிமாற்றத்திற்கு உதவிபுரியும் செயல்திறன் கோட்பாட்டை வேண்டுவர். செயற்கை அறிதிறனில் கருத்துரு வரைபடம் பரவலாகப் பயன்படுத்தப்படுகின்றது. பின்வரும் கருத்துரு வரைபடம் எடுத்துக்காட்டாகத் தரப்பட்டுள்ளது.

பையன் நாயைக் கடித்தான்.

பின்வரும் கருத்துரு வரைபடம் இந்த வாக்கியத்தின் அடிப்படைப் பொருளை உருப்படுத்தம் செய்யும்.

[பையன்]← (செயலி) ←[கடி] →(செ.பொ) →[நாய்]

பெட்டிகள் கருத்துருக்கள் (செவ்வக அடைப்புக் குறிகளுக்குள் தரப்பட்டுள்ளது) என்றும் வட்டங்கள் (பிறை அடைப்புக்குறிக்குள் தரப்பட்டுள்ளது) கருத்துரு உறவுகள் என்றும் அழைக்கப்படும். கருத்துரு எந்த ஒரு மொழியில் வரும் எந்த இருப்புப் பொருளையும் (அதாவது செயலையோ நிலையையோ) உருப்படுத்தம் செய்யும். கருத்துரு உறவுகள் ஒவ்வொரு இருப்புப் பொருளும் பங்காற்றுகிற பாத்திரத்தைக் குறிக்கும். மேற்கண்ட எடுத்துக்காட்டில் [பையன்] என்பது [கடி] என்ற செயலின் (செயலி)-ஆகச் செயல்படும் கருத்துருவை உருப்படுத்தம் செய்யும். [பையன்] என்ற கருத்துரு [கடி] என்பதன் (செயப்படுபொருள்)-ஆகப் பங்கெடுக்கும் இருப்புப் பொருளை உருப்படுத்தம் செய்யும். கருத்துரு வரைபடத்தார் எல்லா பொருண்மை குறித்த செய்திகளையும் ஒரு அடி அமைப்பில் இருப்பதாகக் கொள்ளாமல் ஆறு வேறுபட்ட வகையிலான செய்திகளிலிருந்து வாக்கியங்கள் ஆக்கப்படுவதாகக் கொள்கின்றனர். அவையாவன: இறந்தகாலமும் உள்ளது கிளத்தல் வினை நோக்கும் (indicative mood) செயல் குறிப்பிட்ட நேரத்திற்கு முன்னர் நிகழ்ந்ததைக் குறிப்பிடும். *பையன்* ஒரு சூழலில் ஒரு குறிப்பிட்ட நபரின் முன்வரு கிளவியாக வருகிறது என்பது முற்கோளாகும். *அவன்* சிறிது முன்னர் குறிப்பிடப்பட்டிருக்கலாம் அல்லது அந்த காட்சியில் இருக்கலாம். *நாய்* கடிப்பதைப் பற்றிய புதிய செய்தி கவனக்குவிப்பாகும். பையன் முன்னர் கூறப்பட்ட ஒருவரைக் குறிப்பதன் காரணமாக ஒரு உடனிலைக் குறிப்பு முன்னர் வரும் பையனை *பையன்* என்ற கருத்துருவுடன் தொடர்புபடுத்துகின்றது. ஆறாவதான உணர்ச்சி உணர்பொருள்கள் வாக்கியத்தில் உள்ள வார்த்தைகளாலோ வரைபடத்திலுள்ள கருத்துருக்களாலோ நேரடியாக வெளிப்படுத்தப்படவில்லை. ஆனால் அது மக்களுக்குள் வேறுபடும் முந்தைய அனுபவங்களின் சேர்க்கைகளால் நிறுவப்பெறும்.

9.4.2.2 கருத்துரு வரைபட அகராதியைப் பயன்படுத்தி பொருண்மை மயக்கநீக்கம் செய்தல்

இவ்வாய்வில் மொழியியல் தகவல்களைப் பயன்படுத்தி சொற்பொருண்மை மயக்கநீக்க செய்வதற்கான முயற்சி எடுக்கப்பட்டுள்ளது. இதற்கென்று கருத்துரு வரைபட அகராதி மாதிரியாக உருவாக்கப்பட்டுள்ளது. சொற்கள் அவை உணர்த்தும் பொருண்மை அடிப்படையில் கருத்துருக்களாகப் பிரிக்கப்படும். ஒரு கருத்துரு வரைபடத்தில் அக்கருத்துரு குறித்த எல்லா தகல்களும் வெளிப்படுத்தப்பட்டிருக்கும். இதற்கு புஸ்தெஜொவ்ஸ்கியின் ஆக்கமுறை அகராதி கோட்பாட்டிலிருந்து சில செய்திகள் இங்கு பயன்படுத்தப்பட்டுள்ளன. கருத்துருவரைபட அகராதி கருத்துருக்களுக்கான விளக்கம் மட்டுமன்றி அக்கருத்துகுறித்த எல்லாச் செய்திகளையும் தரும். ஒரு விதத்தில் கூறினால் இது ஒரு களஞ்சியச் செய்திகளை உள்ளடக்கும்.

புஸ்தொஸ்கி கோட்பாடு அடிபடையில் ஒரு சொலுக்குச் பின்வரும் சொற்பொருண்மை தகவல்கள் தரப்படவேண்டும்.

- உறுப்பு உறவு (constitutive relation): ஒரு பொருளுக்கும் (object) அதன் உறுப்புப் பாகங்களுக்கும் (constituent parts) இடையில் உள்ள உறவு.
- வடிவ/முறை உறவு (formal relation): பொருளை (object) பொரிய சொற்களத்திற்குள் வேறுபடுத்தம் செய்கிறது.
- நோக்க உறவு (telic relation): பொருளின் தேவையையும் செயல்பாட்டையும் தருகிறது.
- செயலி உறவு (agentive relation): பொருளின் மூலம் அல்லது அதைக் கொண்டு வருகையில் உட்படும் காரணிகளைத் தருகின்றது.

எடுத்துக்காட்டாகப் புதினம் என்ற கருத்துருவை எடுத்துக்கொள்வோம். அதன் சொற்பொருண்மைத் தகவல்கள் பின்வருவனவற்றை உள்ளடக்கும்.

புதினம்	உறுப்புப்பங்களிப்பு=கதை	.....
	வடிவப் பங்களிப்பு=புத்தகம்	
குண அமைப்பு =	நோக்கப் பங்களிப்பு=படிப்பது	
	செயலிப் பங்களிப்பு=எழுவது	

புத்தகம் என்ற கருத்துருவை எடுத்துக்கொண்டால் அது குறைந்தது பின்வரும் செய்திகளால் விளக்கப்படவேண்டும்.

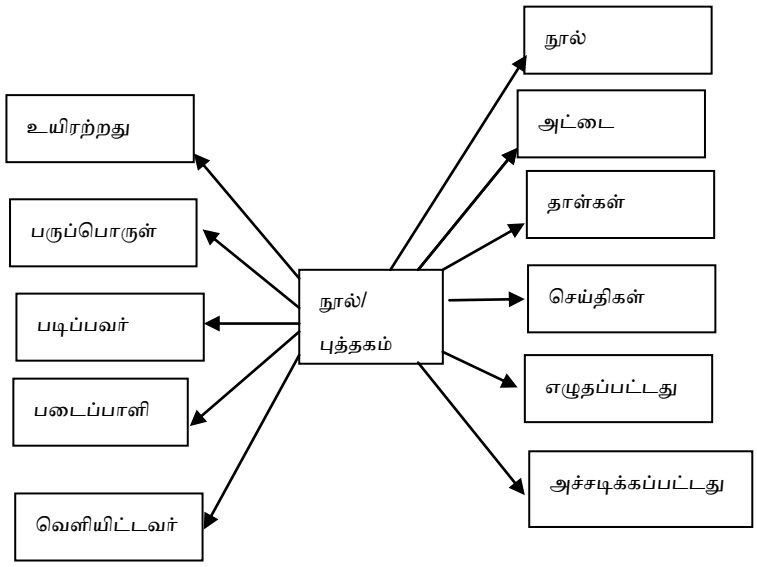
அதன் உறுப்புப் பங்களிப்பு = அது படிப்பதற்கான ஒரு பொருள்

அதன் வடிவப் பங்களிப்பு = அது தாள்களால் ஆனது, அட்டை போடப்பட்டது, படிப்பதற்கான செய்திகள் அடங்கியது

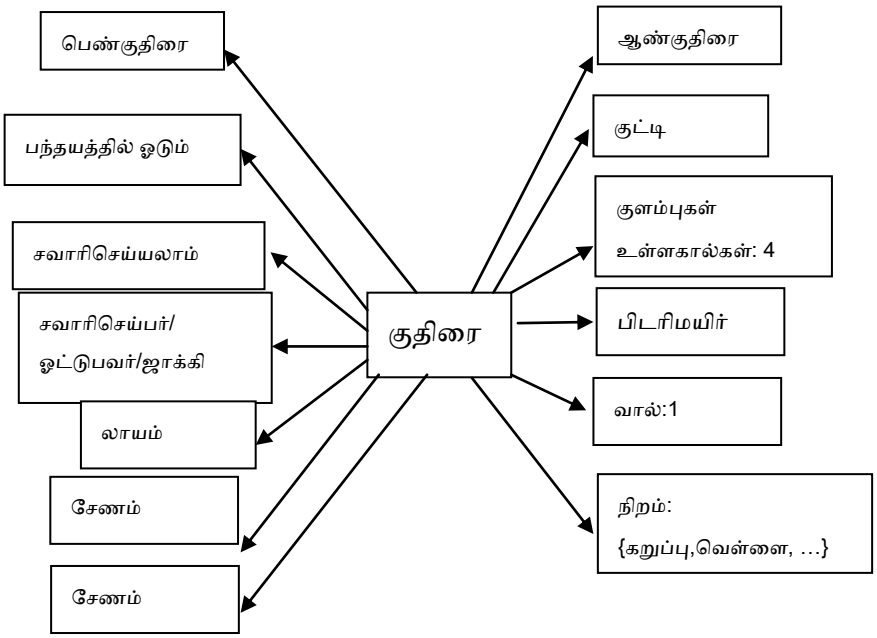
நோக்கப் பங்களிப்பு = படிப்பது

செயலி பங்களிப்பு = எழுதுவது

இதன் படி *புத்தகம்* என்பது பின்வரும் விளக்கத்தைப் பெறும்: படிப்பதற்குப் பயன்படும் எழுதப்பட்ட, வெளியிடப்பட்ட தாளகளால் ஆன, அட்டை போடப்பட்ட, படிப்பதற்கான செய்திகள் அடங்கிய ஒரு திடப்பொருள். இதைப் பின்வருமாறு கருத்துரு வரைபடமாகத் தரல்லாம். கருத்துருக்களுக்கு இடையே உள்ள உறவுகள்தரப்படவில்லை.



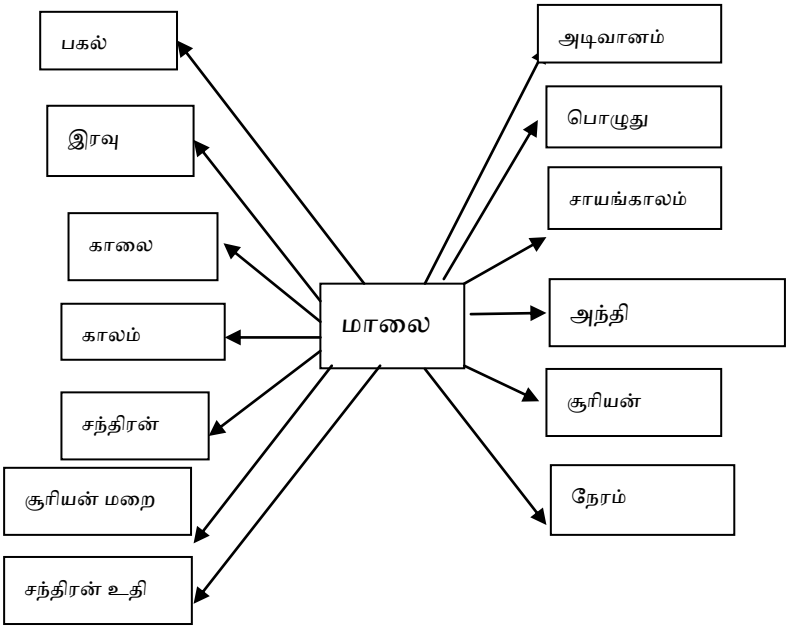
இதுபோல் *குதிரை* என்ற கருத்துருவை எடுத்துக்கொண்டால் அது பின்வருமாறு தகவல்களை வேண்டும்



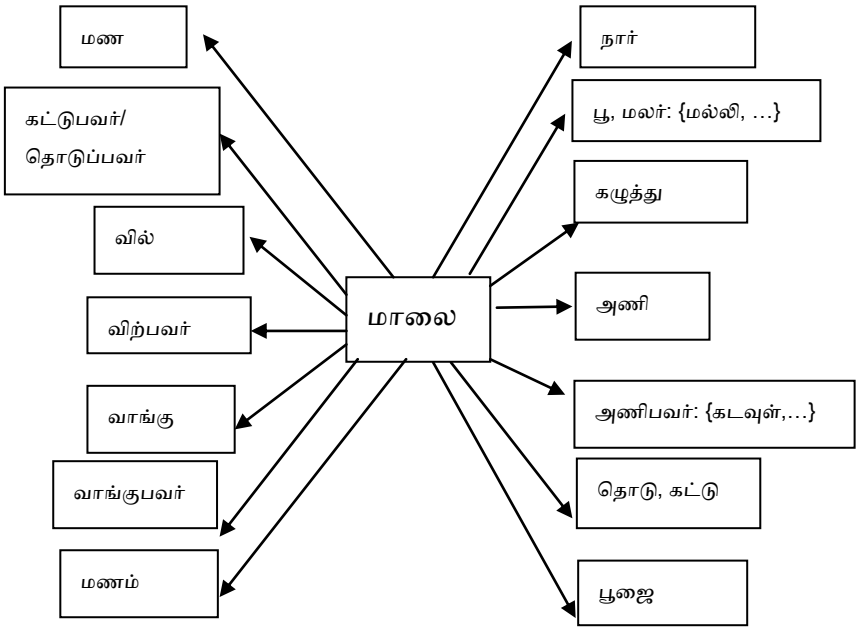
இங்கு முன்மொழிகின்ற கருத்துரு வரைபட அகராதி குறைந்தது மேற்கண்ட தகவல்களைக் கொண்டிருக்க வேண்டும்.

இக்கருத்துரு வரைபட அகராதியைக் கொண்டு முந்தைய அணுகுமுறையில் சுட்டிக்காட்டிய சொற்பொருள் மயக்கம் உள்ள சொற்களின் பொருண்மை மயக்கத்தையும் பிற சொற்களின் பொருள்மயக்கங்களையும் நீக்கலாம். எடுத்துக்காட்டாக *மாலை1* (பொழுது), (கழுத்தில் போடும்) *மாலை2* என்ற கருத்துருவின் கருத்துரு வரைபட அகராதி ஓரளவுகுப் பின்வருமாறு அமையும்.

மாலை1


பகல்
இரவு
காலை
காலம்
சந்திரன்
சூரியன் மறை
சந்திரன் உதி
மாலை
அடிவானம்
பொழுது
சாயங்காலம்
அந்தி
சூரியன்
நேரம்

மாலை2



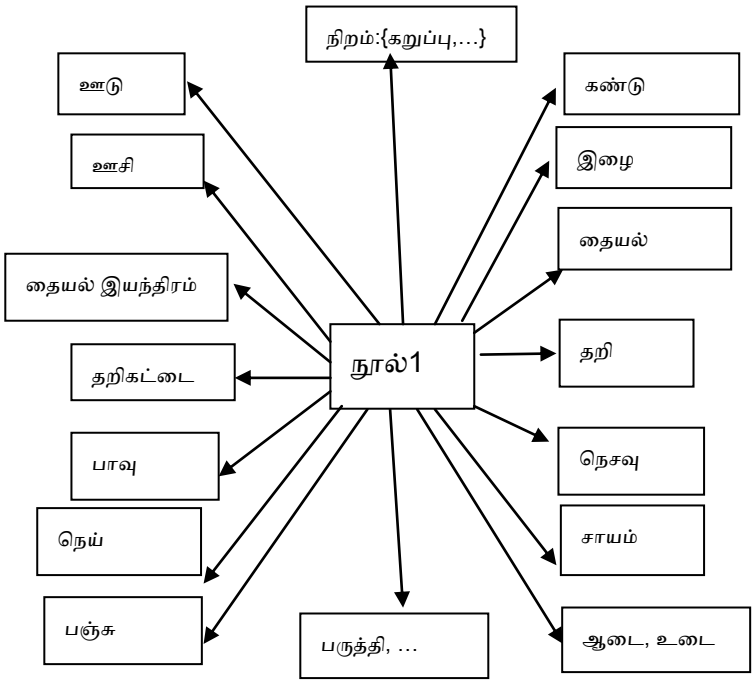
மேற்கண்ட *மாலை1*, *மாலை2* என்ற கருத்துரு வரைபட அகராதி விளக்கம் அடிப்படையில் தரவை பின்வருமாறு பொருண்மை மயக்க நீக்கம் செயலாம்.

எண்	**வாக்கியம்**	**கருத்துரு**
1	தெரசாவின் அறப்பணிகளுக்கு நிகரில்லாத புகழ் மாலைகளாக அமைந்து மகிழ்வித்தன.	மாலை2
2	…தினந்தோறும் காலை மாலை ஆகிய இரு வேளைகளிலும்	மாலை1
3	காலையில் சூரியன் தோன்றிய நேரத்தில் இருந்து மாலையில் மறையும் நேரம் வரை	மாலை1
4	சூரியகாந்திப்பூ காலை, மாலை வேளைகளில் திசை	மாலை1
5	சுற்றுலாப் பேருந்து அடுத்த வெள்ளிக்கிழமை மாலை சென்னைக்குத் திரும்புகிறது	மாலை1
6	காலையில் 5 மணிக்கும், பிறகு 9:30 மணிக்கும், மலையில் 6:30 மணிக்கும் S1	மாலை1

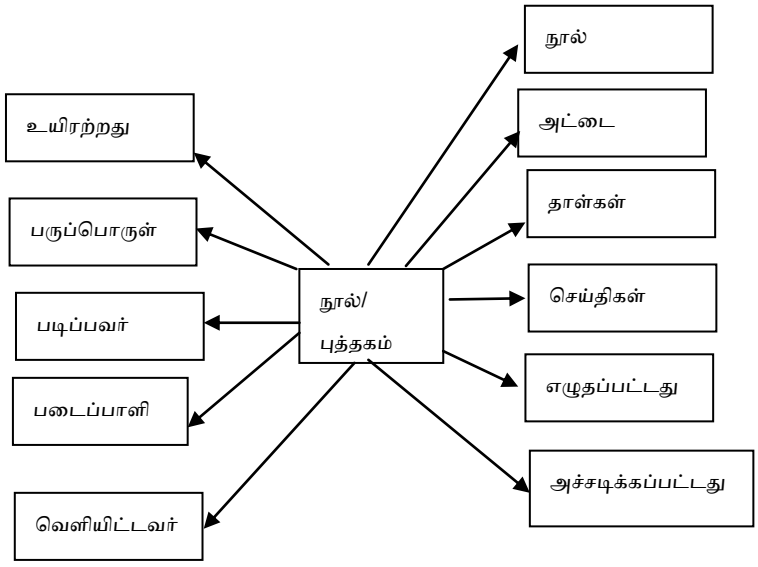
7	அரசு தூக்கிலிட முடிவு செய்த நாளின் முதல் நாள் மலையே, தூக்கிலிட்டுக் S1	மாலை1
8	தத்தும் வீசிய துண்டறிக்கைகள் அன்று மாலையே “இந்துஸ்தான் டைம்ஸ்” S1	மாலை1
9	நடவடிக்கைகளின் காரணமாக நேற்று மாலையிலேயே வந்த செய்தியின்படி S1	மாலை1
10	“குரங்கு கையிலே பூமாலை” என்று அதைப்போல ஊராட்சி S2	மாலை2
11	சிலர், ‘நான் மாலையையே மலர் ஆக்குவேன்	மாலை2
12	அழகு, இளமை, கருவிழி, செல்லம், வாகனம், மாலை, கொடி முதலியவை	(மாலை2)
13	அல்லது திருநீறு காணப்படும் சபையோர் போடும் மாலைகளைப் கழற்றாமலே	மாலை2
14	நானும் அந்தச் செய்தியை நேற்று மாலை அமைச்சரவைக் கூட்டம்	மாலை1
15	உணவுப் பொருள்களை உட்கொள்ளும் பொழுதும், மாலையில் மேய்ச்சலுக்குப் பிறகும்	மாலை1

முந்தைய அணுகுமுறையில் கூறப்பட்ட பொருண்மை மயக்கம் உள்ள *நூல்1* (இழை), *நூல்2* (புத்தகம்) என்ற கருத்துருக்களைக் கருத்துரு வரைபட அகராதி அடிப்படையில் பொருண்மை மயக்கம் நீக்குவதைக் கருதுவோம். *நூல்2இன்* கருத்துரு வரைபட அகராதி முன்னர் தரப்பட்டுள்ளது; இது மீண்டும் இங்கு ஒப்புமைக்காகக் காட்டப்பட்படுள்ளது. *நூல்1*-இன் கருத்துரு வரைபட அகராதி பின்வருமாறு அமையும்.

நூல்1


நிறம்:{கறுப்பு,…}
ஊடு
ஊசி
தையல் இயந்திரம்
தறிகட்டை
பாவு
நெய்
பஞ்சு
நூல்1
பருத்தி, …
கண்டு
இழை
தையல்
தறி
நெசவு
சாயம்
ஆடை, உடை

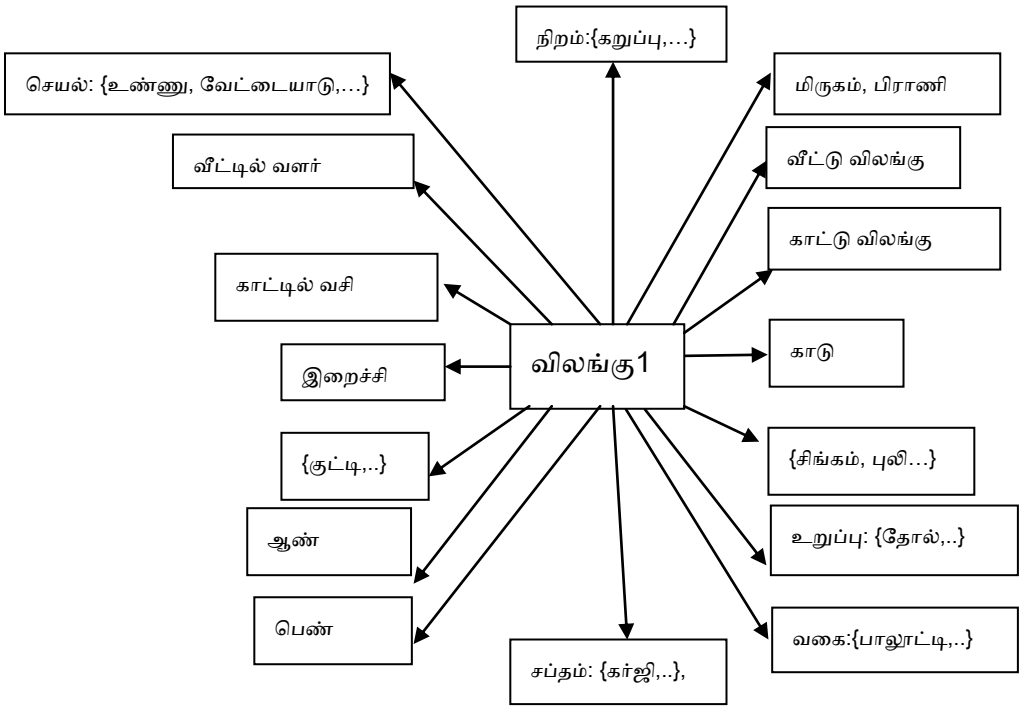
நூல்2


உயிரற்றது
பருப்பொருள்
படிப்பவர்
படைப்பாளி
வெளியிட்டவர்
நூல்/
புத்தகம்
நூல்
அட்டை
தாள்கள்
செய்திகள்
எழுதப்பட்டது
அச்சடிக்கப்பட்டது

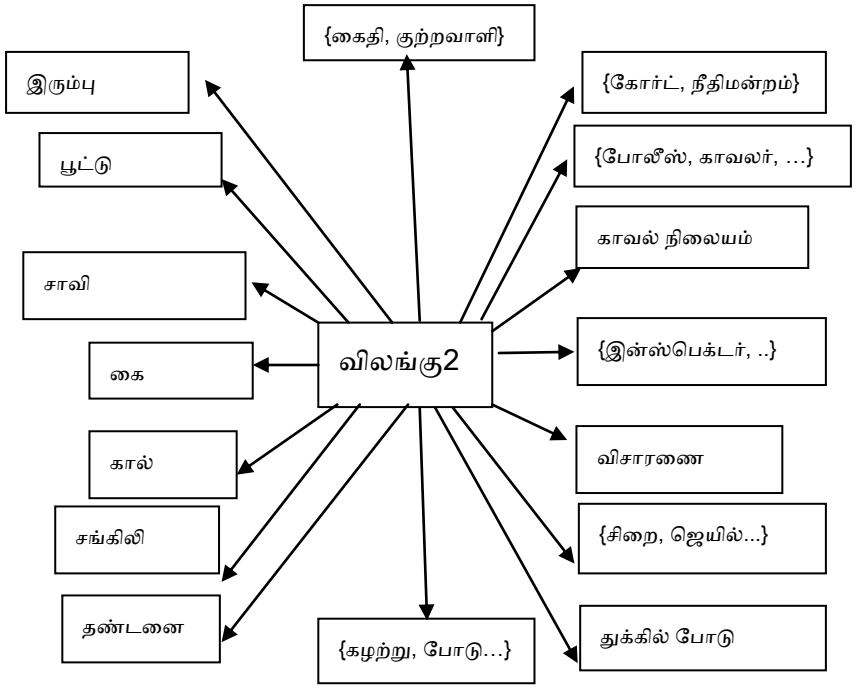
எண்	வாக்கியம்	கருத்துரு
1	… சட்டையின் நூல் இழை….	நூல்1
2	தறியில் நூல் சிக்கி துணி நெய்ய..	நூல்1
3	…நூலில் சாயம் ஏற்றி…	நூல்1
4	கதர் துணி நெய்ய நூல் தேவை…	நூல்1
5	தறியில் ஊடு…நூல்	நூல்1
6	ஊசியில் நூல் கோர்க்க இயல…	நூல்1
7	நூல் ஆசிரியர் பெயர் தெரிவில்லை	நூல்2
8	பழனியப்பா நூல் பதிப்பகத்தில் அச்சடிக்கபட்ட…	நூல்2
9	நூல் 1000 பக்கங்களைக் கொண்டு ….	நூல்2
10	நூல் 1/8 டெம்மி அளவு தாள் கொண்டு …	நூல்2

அடுத்ததாகப் முந்தைய அணுகுமுறையில் பொருண்மை மயக்கம் உள்ளவை என்று எடுத்துக்கொண்ட *விலங்கு* (மிருகம்), *விலங்கு* (கைவிலங்கு) என்ற கருத்துருக்களைக் கருத்துரு வரைபட அகராதி அடிப்படையில் பொருண்மை மயக்கம் நீக்குவதைக் கருதுவோம். *விலங்கு1*, *விலங்கு2* இவற்றின் கருத்துரு வரைபட அகராதி பின்வருமாறு அமையும்.

விலங்கு1


நிறம்:{கறுப்பு,…}
செயல்: {உண்ணு, வேட்டையாடு,…}
மிருகம், பிராணி
வீட்டில் வளர்
வீட்டு விலங்கு
காட்டு விலங்கு
காட்டில் வசி
விலங்கு1
இறைச்சி
காடு
{குட்டி,..}
{சிங்கம், புலி…}
ஆண்
உறுப்பு: {தோல்,..}
பெண்
வகை:{பாலூட்டி,..}
சப்தம்: {கர்ஜி,..},

விலங்கு2


{கைதி, குற்றவாளி}
இரும்பு
{கோர்ட், நீதிமன்றம்}
பூட்டு
{போலீஸ், காவலர், …}
காவல் நிலையம்
சாவி
விலங்கு2
{இன்ஸ்பெக்டர், ..}
கை
விசாரணை
கால்
{சிறை, ஜெயில்...}
சங்கிலி
தண்டனை
துக்கில் போடு
{கழற்று, போடு…}

எண்	தரவு வாக்கியங்கள்	கருத்துரு
1	விலங்குகள் காட்டில் அலைந்து திரிந்து…	விலங்கு1
2	காட்டு விலங்கான சிங்கம்…	விலங்கு1
3	காட்டில் சிங்கம், புலி, யானை,…ஆகிய விலங்குகள்..	விலங்கு1
4	விலங்குகளில் பெரிதான யானையின் தந்தங்கள் விலைமிக்கவை என…	விலங்கு1
5	நாய், பூனை, பசு, குதிரை போன்ற வீட்டு விலங்குகளை…	விலங்கு1
6	போலீஸ் குற்றவாளிகளைக் கைது செய்ய…	விலங்கு2
7	கைதியை கோர்ட்டில் கொண்டுபோக…	விலங்கு2
8	கைதிகளை விசாரணை செய்து தண்டனை அளிக்க…	விலங்கு2

உள்ளடங்கு-உள்ளடக்கு மொழி உறவுகள், எதிர்மொழிகள் உறவுகள், முழு-சினை மொழி உறவுகள், வகை உறவுகள், உறுப்பினர் உறவுகள், சொந்த பந்த உறவுகள், சமூக உறவுகள் போன்ற எல்லா உறவுகளையும் கருத்துரு வரைபட அகராதியாகத் தர இயலும்.

9.4.3. செம்மொழிப் பனுவல்களில் சொற்பொருண்மை மயக்கநீக்கம்

மேற்சொன்ன இரு அணுகுமுறைகளையும் செம்மொழிப் பனுவல்களில் இடம்பெறும் சொற்களின் சொற்பொருண்மை மயக்கங்களை நீக்கப் பயன்படுத்த இயலும். பின்வருவன இதற்கு அறிவு மூலவளங்களாக அமையும்.

1.சங்க இலக்கியப் பொருட் களஞ்சியம். தமிழ்ப் பல்கலைக் கழகம் தஞ்சாவூர்.

2. பெருஞ்சொல் அகராதி. தமிழ்ப் பல்கலைக் கழகம் தஞ்சாவூர்.

4. சங்க இலக்கியச் சொல்லடைவு

5. சங்க இலக்கிய சொல்லாட்சி அகராதி மற்றும் பொருளடைவு

எடுத்துக்காட்டாகப் புறநானூற்றிலிருந்து எடுக்கப்பட்ட *அணி* என்ற சொல்லின் சொற்பொருளடைவின் மாதிரி பின்வருமாறு பத்து பொருண்மைகளை வேறுபடுத்தி அமையும்.

அணி

3. அழகு

அழகு தாழ்ந்த அணி (புறம் 57.1)

3.1 அணிகொள் தேர் (புறம் 68.14)

3.2 ….

4. அணிகலன்

இழையணிபொலிந்த (புற, 81.1)

5. ஒப்பனை

ஆடு கோட்பாட்டுச் சேரலாதன் போந்தைப் பொழியின் கண் அணியல் (படி 51.9)

6. அணிமை
7. வரிசை
8. திரட்சி
9. ஆடை
10. ஒப்புமை
11. சூழ்தல்
12. பக்கரை (சேணம்)

முன்னர் விளக்கிய சேர்ந்துவருகை அணுகுமுறையைப் பயன்படுத்தி *அணி* என்பதன் பத்து பொருண்மைகளையும் வேற்படுத்தி அறியலாம்.

சங்க இலக்கியப் பொருட்களஞ்சியத்தைப் பயன்படுத்தியும் மேற்சொன்ன அணுகுமுறை அடிப்படையில் சொற்களின் பொருண்மை மயக்கங்களை நீக்கலாம். எடுத்துக்காட்டாக, *களவு* என்ற சொல்லை எடுத்துக்கொண்டால், சங்க இலக்கியப் பொருட்களஞ்சியம் அதற்கு மூன்று பொருண்மைகளை வேறுபடுத்தி அவை வரும் இடங்களையும் பட்டியலிட்டுள்ளது.

1. களவொழுக்கம்

தித்தனின் உறையூரைச் சூழ்ந்த புறங்காடு போல நம் களவும் பல இடையூறுகளைக் கொண்டது (அகம்122:21-23)

நள்ளிரவில் தலைவனுடன் புணர்குறி வாய்த்த களவும் எல்லை கடந்து நீண்டுவிட்டது (அகம் 325:3,4)

….

2. திருட்டு

கொலைத்தொழிலை நீக்கியும் களவை நீக்கியும் உழவு வாழ்க்கையை விரும்பிய வேளாண்மை மாந்தர் வாழ்ந்தனர். (பட்.199, 205)

களவாகிய உழவினைச் செய்யும் கடுங்கண் மழவர் தம் தொழிலுக்கு ஒடுங்காடு செல்வர்.(அகம் 91:11,12).

….

3. களாச்செடி

களவு புளிப்புப்போல விளைந்த கள் புளிப்பேறியிருந்தது (புறம் 328:8)

களவும் பழுத்து நைந்து புளிப்பேறின (அகம் 394:1)

9.5.முடிவுரை

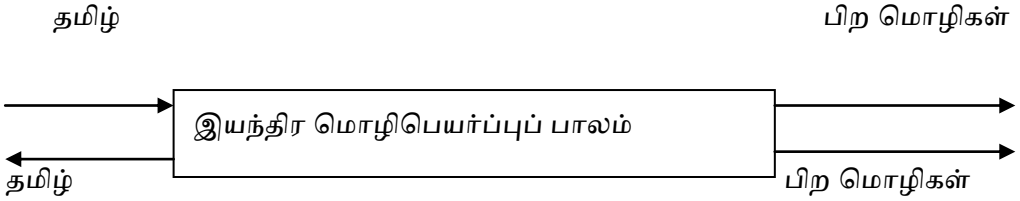
சொற்பொருண்மை மயக்கநீக்கத்திற்குப் பல வழிமுறைகள் உள்ளன. கண்காணிக்கபட்ட அணுகுமுறைகளைக் காட்டிலும் கண்காணிக்கப்படாத அணுகுமுறைகள் சிக்கனமானது. அறிவு அடிப்படையிலான அணுகுமுறைகள் வரவேற்கப்படவேண்டியவை ஆகும். மொழியியல் அடிப்படையில், குறிப்பாகச் சொற்பொருண்மையியல் அடிப்படையில் பொருண்மைமயக்கச் சிக்கலுக்குத் தீர்வுகாண முயல்வேண்டும். பொருண்மைமயக்கச் சிக்கல் சொற்பொருண்மை மொழியியல் அடிட்படையில் அணுகப்பட்டால் தான் தீர்வுகாண்பதற்கான வழிமுறைகளை நிறைவாக/தீர்மானமாகக் காண இயலும். கொத்து/திரட்டு நெறிமுறை கண்காணிக்கப்படாத அணுகு முறைகளாகும். இம்முறை ஓரளவுக்குப் பலன் விளைவிப்பதாகும். அவர்தம் வேற்றுமைக் குறியீட்டு முறை அவ்வளவு நம்பத்தகுந்தது அல்ல. தற்போது சொல்வலைகளைப் பயன்படுத்தி பொருண்மை மயக்கநீக்கம் செய்யும் முயற்சிகள் நடந்து வருகின்றன. கருத்துரு வரைபடம் அடிப்படையிலான அணுகுமுறை மொழியில் மற்றும் சொற்பொருண்மையியல் கோட்பாடுகளின் அடிப்படையில் அமைகிறது. மேற்சொன்ன கருத்துரு வரைபட அகராதியை உருவாக்குவது சவாலுக்கு உரியதாகும். இம்முறை சிக்கமானதல்ல. இருப்பினும் இந்நெறிமுறை வரவேற்கத்தகுந்தாகும். கண்காணிக்கப்படாத அணுகுமுறைகளை அல்லது புள்ளியியல்சார் அணுகுமுறைகளைப் பின்பற்றும் முன்னர் அறிவு அடிப்படையிலான சொற்பொருண்மையியல் சார் அணுகுமுறைகளுக்கான ஆயத்தங்கள் மேற்கொள்ளப்படவேண்டும். ஒரு கலப்படமான, அதாவது கண்காணிக்கப்பட்ட, கண்காணிக்கப்படாத மற்றும் அறிவு அடிப்படையிலான அணுகுமுறைகளை இணைத்துச் செய்யப்படும் அணுகுமுறை நம்பத்தகுந்ததாயும் தீர்வுகாண்பதாயும் சிறப்பாகவும் அமைய இயலும். தமிழில் சொற்பொருண்மை மயக்கநீக்க இராசேந்திரனால் செய்யப்பட்டு நல்ல விளைவை எய்தியுள்ளது (Rajendran et al 2013).

இயல் 10

இயந்திர மொழிபெயர்ப்பு

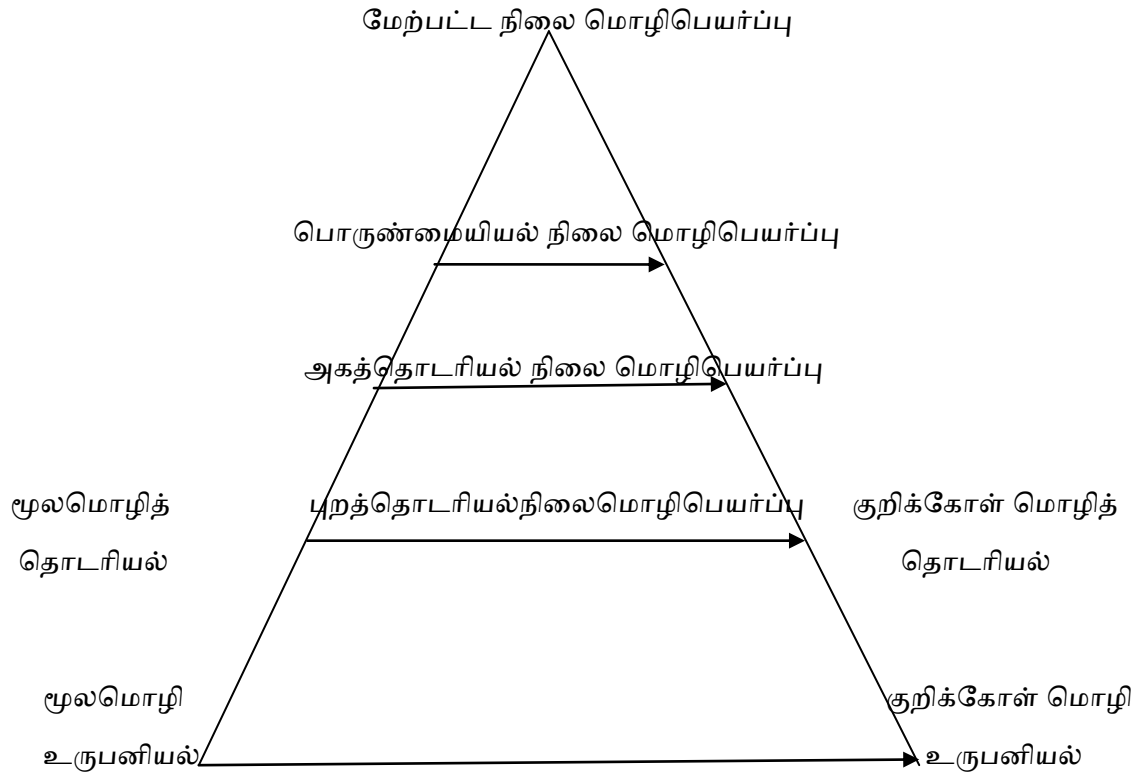
10.1. அறிமுகம்

ஒரு மொழியிலுள்ள செய்திகளைப் பிற மொழிக்கு மாற்றித்தருவதை மொழிபெயர்ப்பு என்று சுருக்கமாய்க் கூறலாம். மொழி பெயர்ப்பு செய்வதற்கு மனித மூளைதான் தலைசிறந்த கருவி என்றாலும், மனித மொழிபெயர்ப்பில் ஏற்படும் கால தாமதத்தைத் தவிர்க்க இயந்திர மொழிபெயர்ப்பைக் கைக்கொள்ளலாம். இயந்திர மொழிபெயர்ப்பு செய்வதற்கான இயந்திரம் கணிப்பொறிதான். தமிழிலிருந்து பிற மொழிகளுக்கு மொழிபெய்ர்ப்பதற்கும் பிற மொழிகளிலிருந்து தமிழ்மொழிக்கு மொழிபெயர்ப்பதற்கும் ஒரு இயந்திர மொழிபெயர்ப்புப் பாலம் அமைத்தால் அது மொழிப் பரிமாற்றத் தடங்கலை உடைத்தெறியும்.



மூல மொழி மற்றும் குறிக்கோள் மொழி இவற்றின் கட்டமைப்புக்குத் தகுந்தவாறு இயந்திர மொழிபெயர்ப்பின் தன்மை மாறும். அதிக ஒற்றுமையுள்ள இரு மொழிகளைச் சொல் அல்லது உருபனியல் நிலையிலேயே நேரடியாக மொழி பெயர்க்கலாம். மொழிகளின் கட்டமைப்பு வேறுபாட்டிற்குத் தகுந்தவாறு, புறத்தொடர் நிலையிலோ அகத்தொடரியல் நிலையிலோ பொருண்மையியல் நிலையிலோ மொழிபெயர்க்கலாம்.

எந்த அளவிற்கு இயந்திர மொழிபெயர்ப்பில் மனித உழைப்புப் பயன்படுகிறது என்பதன் அடிப்படையில் இயந்திர மொழிபெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்.



எந்த அளவிற்கு இயந்திர மொழிபெயர்ப்பில் மனித உழைப்பு பயன்படுகிறது என்பதன் அடிப்படையில் இயந்திர மொழி பெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்.

1. மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பு
2. கணிப்பொறி உதவியுடன் நடைபெறும் இயந்திர மொழி பெயர்ப்பு
3. கணிப்பொறி தானே மொழிபெயர்க்கும் இயந்திர மொழி பெயர்ப்பு

இயந்திர மொழிபெயர்ப்புக்குப் பொருண்மையியல் ஆய்வி மற்றும் ஆக்கி, தொடரியல் ஆய்வி மற்றும் ஆக்கி, மொழிபெயர்ப்பு அகராதி என்பன போன்ற துணைக்கருவிகளைத் தயாரித்தல் வேண்டும்.

10.2. கணிப்பொறி வழி மொழி பெயர்ப்பின் தேவை

ஒரு மொழியில் உள்ள தகவலை அதன் கருத்து மாறாமல் மற்றொரு மொழியில் மாற்றுவதையே மொழிபெயர்ப்பு என்கிறோம். இத்தகைய மொழி பெயர்ப்பில் தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அந்த மொழியின் இயல்புத் தன்மை மாறாது பாதுகாத்தல் வேண்டும். தகவல் எந்த மொழியில் உள்ளதோ அதை மூலமொழி (Source Language) என்கிறோம். தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகின்றதோ அந்த மொழியை

இலக்குமொழி (Target Language) என்கிறோம். இத்தகைய மொழி பெயர்ப்புகளை மனித முயற்சியால் மட்டும் மேற்கொள்வது என்பது கடினமானப் பணியாகும்.

எனவே மொழிபெயர்ப்புப் பணிகளைக் கணிப்பொறி வழி மேற்கொள்வதற்கான கட்டாயம் ஏற்படுள்ளது. கணிப்பொறி வழி மொழிபெயர்ப்பின் தேவைகளும் கட்டாயங்களும் கீழேப் பட்டியலிடப்பட்டுள்ளன.

- உலகளவில் தகவல்கள் அன்றாடம் மலையெனக் குவிகின்றன. அவைகளை உடனுக்குடன் மொழிபெயர்த்தல் அவசியமாகிறது.
- கணிப்பொறி வழி மொழிபெயர்ப்பு மனித உழைப்பைக் குறைக்கிறது.
- குறைந்த நேரத்தில் நிறைய மொழி பெயர்ப்புப் பணிகளைச் செய்ய இயலுகிறது. இதனால் காலவிரயம் தவிர்க்கப்படுகிறது.
- தற்போதைய மொழி பெயர்ப்புப் பணிகளுக்கு ஏற்ற வகையில் மொழிப் பெயர்ப்பாளர்கள் இல்லை.
- மனித முயற்சியில் மொழிபெயர்கின்ற பொழுது மொழி பெயர்ப்பாளர்கள் மூலமொழியிலும் இலக்கு மொழியிலும் புலமை உள்ளவர்களிடம் கருத்துக்களைப் பெற்று அதை வழியமைப்பு மொழயில் எழுதி மொழிபெயர்ப்பு செய்யலாம்.

10.3. கணிப்பொறி வழி மொழிபெயர்ப்பின் வகைகள்

மொழிபெயர்ப்பில் கணிப்பொறி எவ்விதமாகப் பயன்படுத்தப்படுகின்றது என்பதன் அடிப்படையில் மொழிபெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்:

1. கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு
   (CAMT – Computer Assisted Human Translation)
2. மனித உதவியின் கணிப்பொறி மொழிபெயர்ப்பு
   (HAMT-Human Assisted Machine Translation)
3. முற்றிலும் தானியங்கி மொழிபெயர்ப்பு
   (Automatic Machine Translation)

கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு

இதில் கணிப்பொறி உதவியுடன் நடைபெறும் மனிதமொழி பெயர்ப்பில் மனித மொழிப் பெயர்ப்பாளர்களுக்கு முக்கியத்துவம் தரப்படுகின்றது. அவர்கள் மொழிப்பெயர்ப்பப்புப் பணியில் ஈடுபடும் பொழுது ஒரு மொழியில் உள்ள சொற்களுக்கு நிகரான பிறமொழிச் சொற்களைத் தெரிந்துகொள்ளவும் அச்சொற்களின் பொருள் பயன்பாடு ஆகியவற்றை எளிமையாகத் தெரிந்துகொள்ளவும் கணிப்பொறி பயன்படுத்தப்படுகிறது. மேலும் ஒரு மொழியின் வாக்கிய அமைப்பு, தொடரமைப்பு போன்றவைகளுக்கு நிகரான பிறமொழி அமைப்புகளை விரைவாகத்

தயார் நிலையில் பெறுவதற்கும் கணிப்பொறி உதவுகிறது. இவ்வகை மொழிபெயர்ப்பில் மொழிப்பெயர்பாளர்கள் தான் முதல் நிலையில் செயலாற்றுவர்; அவர்களுக்கு உதவும் நிலையில் கணிப்பொறி செயலாற்றும்.

மனித உதவியுடன் கணிப்பொறி மொழிப்பெயர்ப்பு

மனித உதவியுடன் நடைபெறும் இயந்திர மொழிப்பெயர்ப்பில் கணிப்பொறி முதன்மை நிலையை வகிக்கிறது. கணிப்பொறி மொழிப்பெயர்ப்புப் பணிகளைச் செய்கின்ற பொழுது முன்னர் பதிவு செய்யப்பட்டுள்ள நகலுக்கு மேலாக ஏதாவது செய்திகள் தேவைப்படின் அந்நிலையில் மனித உதவியை நாடுகிறது. மேற்கண்ட மூவகை மொழிப்பெயர்ப்பு வகைகளில் மனித உதவியுடன் செயல்படும் இயந்திர மொழி பெயர்ப்பு தான் நடைமுறையில் பரவலாக உள்ளது.

முற்றிலும் தானியங்கி மொழிபெயர்ப்பு

முற்றிலும் இயந்திர மொழிபெயர்ப்பு என்பது தற்போதைய நிலையில் சாத்தியமானதல்ல இதற்கு எண்ணற்ற மொழியில் செயல்பாடுகளும் மொழிபெயர்ப்பு முறைமைகளும் வடிவமைக்கப்பட்ட வேண்டி உள்ளது. இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்பில் முற்றிலும் தானியங்கப்படுத்தும் முயற்சி இலக்காகக் கொள்ளப்பட்டுள்ளது.

10.4.கணிப்பொறி வழி மொழிப்பெயர்ப்பின் சில முயற்சிகள்

கணிப்பொறி வழி மொழி பெயர்ப்பினைத் தானியங்கி மொழிப்பெயர்ப்பு (Automatic Translation) என்றும் இயந்திர மொழிபெயர்ப்பு (Machine Translation) என்றும் கணிப்பொறி உதவியுடன் நடைபெறும் மொழிபெயர்ப்பு என்றும் கூறலாம். கணிப்பொறியின் வேகம், நினைவகத்திறன், பதிவு செய்துள்ள செய்திகளைத் தேவையான பொழுது தேவையான வடிவமைப்பில் மிக விரைவாக மீளப்பெறக்கூடிய வசதி (Retrievel System) போன்றவைக் கணிப்பொறி வழி மொழி பெயர்ப்புப் பணியை ஊக்குவிக்க உதவின.

கணிப்பொறி வழி மொழிபெயர்ப்புப் பணி 1950-ஆம் ஆண்டு வாரன் வீவர் (Warren Waiver) என்ற அறிஞரால் தொடங்கப்பட்டது. இதைத் தொடர்ந்து பூபார் ஹில்லல் என்ற அறிஞர் கணிப்பொறி வழி மொழிபெயர்ப்புப் பணியைத் தொடங்கினார். 1954-ஆம் ஆண்டு ஜார்ஜ் டவுன் பல்கலைக்கழகம் கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியைத் தொடங்கியது. இதைத் தொடர்ந்து 1956-ஆம் ஆண்டு மாஸ்கோவில் பிரெஞ்சு மொழியிலிருந்து ரஷ்ய மொழிக்குக் கணிப்பொறி வழி மொழிப்பெயர்ப்பு செய்யும் பணி நடைபெற்றது. இப்பணியை மேல் சங் (Mel Chunk) என்ற அறிஞர் மேற்கொண்டார்.

1973-ஆம் ஆண்டு அமெரிக்காவில் உள்ள Logos Incoration Limited என்ற நிறுவனம் ஒரு மென்பொருளை வடிவமைத்து. இம்மென் பொருளைக் கொண்டு ஆங்கிலத்தில் அமைந்துள்ள செய்திகளை வியட்நாம் மொழியில் மொழிப்பெயர்ப்பு செய்தது. ஏறக்குறைய இதே காலக்கட்டத்தில் Tertils என்ற நிறுவனம் TUTUS என்ற மென்பொருளை வடிவமைத்தது. இம்மென் பொருளைக் கொண்டு ஆங்கிலம், ஸ்பானிஷ், பிரெஞ்சு மற்றும் ஜெர்மன் ஆகிய மொழியில் உள்ள செய்திகள் ஒன்றிலிருந்து மற்றவைகளுக்கு மொழிமாற்றம் செய்யப்பட்டது. இதைத் தொடர்ந்து CUTD-Chinese University Translator) என்ற மென்பொருள் உருவாக்கப்பட்டது. இதைக்கொண்டு சீனமொழியில் உள்ள செய்தித்தாள்கள் ஆங்கில மொழிக்கு மொழிமாற்றம் செய்யப்பட்டன.

1974-ஆம் ஆண்டு SYSTRAN என்ற மென்பொருள் உருவாக்கப்பட்டது. SYSTRAN என்பதன் விரிவு System Translation என்பதாகும். இம்மென்பொருள் ரஷ்ய மொழியிலிருந்து பிரஞ்சு மொழிக்கு மொழி மாற்றம் செய்யப்பயன்பட்டது. இதற்கானத் திட்டத்தை P.Toma என்பவர் மேற்கொண்டார். வணிக ரீதியில் முதன்முதலில் வெளிப்யிடப்பட்ட கணிப்பொறி வழி மொழிபெயர்ப்புக்கான மென்பொருள் இதுவேயாகும்.

1975-ஆம் ஆண்டு TAUM என்ற மென்பொருள் உருவாக்கப்பட்டது. இம்மென் பொருள் ரஷ்ய மொழியில் மாற்றம் செய்யப்பட்டது. இதே காலக்கட்டத்தில் GETA என்ற மென்பொருள் ஆங்கில மொழியிலிருந்து பிரெஞ்சு மொழிக்கு வழியமைக்க உருவாக்கப்பட்டது. இது போன்று உலகளாவிய நிலையில் ஆங்காங்கே சோதனைக்கட்ட முயற்சியாலும் வணிக நோக்கத்தாலும் கணிப்பொறி வழி மொழிபெயர்க்க மென்பொருள்களை உருவாக்கும் திட்டங்கள் தொடர்ந்து நடைபெற்று வருகின்றன.

1985-ல் தமிழ்ப்பல்கலைக் கழகத்தில் TUMTS (Tamil University Mechine Translation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது. ரஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழி பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப் பல்கலைக்கழகக் கணிப்பொறி அறிவியல் துறை, மொழியியல் துறை, மொழிப்பெயர்ப்புத்துறை ஆகியவைகளின் கூட்டு முயற்சியில் நடைமுறைப்படுத்தப்பட்டது. இதற்குப்பின்னால் கான்பூரில் உள்ள இந்திய தொழில்நுட்ப நிறுவனம் Indian Institute of Technology இந்தியிலிருந்து ஆங்கிலத்திற்கும் ஆங்கிலத்திலிருந்து இந்திக்கும் மொழிபெயர்ப்பு செய்வதற்கும் இந்திய மொழிகளை இந்திக்கும் இந்தியை பிற இந்திய மொழிகளுக்கும் மொழிபெயர்ப்பதற்கான ஆரம்ப கட்ட வேலைகளை மேற்கொண்டு செயலாற்றியது. அனுசாரகா (*anusaraka*) என்ற இயந்திர மொழிபெயர்ப்பு மென்பொருள் இதன் விளைவாகும்.

10.5. கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்

ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழி மாற்றம் செய்வதற்கான மென்பொருள்கள் பல உருவாக்கப்பட்டு வருகின்றன. பெரும்பாலும் இத்ததைய கணிப்பொறி வழி மொழி பெயர்ப்புப் பணிகளில் எல்லைகள் இரண்டாக வரையறுத்து உள்ளன.

1. மொழி என்பது பல வட்டாரங்களில் மக்கள் பேசும் மற்றும் எழுதும் அமைப்புகளின் தொகுப்பாகும். ஒரு வட்டாரத்தில் பயன்படுத்தக்கூடிய சொற்கள் பிற வட்டாரங்களில் பயன்படுத்தக்கூடிய உறுதியில்லை. தமிழ் போன்ற மொழிகளில் மொழி நடை இலக்கியமொழி, பேச்சுமொழி என இருவகையாகப் பிரிக்கப்படுகிறது. பேச்சுமொழியைக் கணிப் பொறியாக்கம் செய்வது என்பது சுலபமான பணியன்று. எனவே, மொழிப்பெயர்ப்புப் பணியில் இலக்கிய மொழிகள் மட்டுமே எடுத்துக்கொள்ளப்படுகின்றன.
2. செய்யுள், கவிதை போன்றவைகளில் பொருளுக்குத் தரும் முக்கியதுவத்தைவிட கருத்தைப் புலப்படுத்தும் உத்தி, உணர்ச்சி, வேகம் போன்ற கூறுகளுக்கு முக்கியத்துவம் தரப்படுகின்றன. இவைகள் மனித உளவியலைப் பிரதிப்பலிப்பன. இவ்வுளவியல் கூறுகளைக் கணிப்பொறி வழி வெளிப்படுத்துவது என்பது இயலாத ஒன்றாகும். இதேபோல் இலக்கியங்களிலும் பொருளைவிட உளவியல் கூறுகளுக்கு தான் முக்கியத்துவம் தரப்படுகின்றது. எனவே, செய்யுள்கள், கவிதைகள், பாடல்கள் மற்றும் இலக்கியச் செய்திகள் ஆகியவைகளைக் கணிப்பொறி வழி மொழிபெயர்ப்பதற்குத் தொடக்கநிலையில் பயிற்சி எடுத்துக்கொள்ளப்படவில்லை. அறிவியல் தொழில்நுட்ப சார்ந்த படைப்புகளில் பொருள்களுக்கே முக்கியத்துவம் தரப்படுகிறது. இங்கு கோபம், வீரம், தாபம், மகிழ்ச்சி போன்ற உணர்வுகளுக்கு முக்கியத்துவம் தரப்படுவதில்லை. எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணிகளுக்குப் பொதுவாக அறிவியல் மற்றும் தொழில்நுட்பத்தைச் சார்ந்த படைப்புகளே எடுத்துக்கொள்ளப்படுகின்றன.

இதேப் போன்று மொழிகளின் தன்மையைப் பொறுத்துக் கணிப்பொறி மொழிபெயர்ப்பினை இருபெரும் பிரிவுகளாகப் பிரிக்கலாம்:

1. இருதொடர்புடைய மொழிகளுக்கிடையே அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Interlingual Method of Machine Translation): இலக்கண அமைப்பிலும் பொருள் அமைப்பிலும் தொடர்பும் நெருக்கமும் உடைய இருமொழிகளுக்கிடையே செய்யப்படும் கணிப்பொறி மொழிப்பெயர்ப்பு இவ்வகையைச் சார்ந்தாகும்.
2. இடைநிலை மொழி அடிப்படையில் அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Intermediate Language Based Machine Translation) இலக்கணம் மற்றும்

மொழியமைப்பில் முற்றிலும் வேறுபட்ட இருமொழிகளுக்கிடையே கணிப்பொறி வழி மொழிபெயர்ப்பை நடைமுறைப்படுத்த இம்முறை உதவுகிறது. இம்முறையில் மூலமொழியில் உள்ள சொற்றொடர்கள் இடைநிலை மொழி வழி இலக்கு மொழிக்கு மாற்றம் செய்யப்படுகின்றது.

10.6. கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

மொழிபெயர்ப்புப் பணியைக் கணிப்பொறி வழி மேற்கொள்கின்ற பொழுது கணிப்பொறியின் சிறப்புத் தன்மை, துல்லியம், வேகம், நினைவகக் கொள்திறன், தகவல் தளம் உருவாக்கும் வசதிகள் போன்றவை மொழிபெயர்ப்புத் திறனை நிர்ணயம் செய்யும். ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழிபெயர்ப்பு செய்யும் போது பன்மொழித் திறமை அவசியம் என்ற நிலை இல்லை. அந்தந்த மொழியின் வல்லுநர்களின் மொழித்திறன் கணிப்பொறி வழியமைப்பாக வழியமைக்கப்படுகிறது. மனித மொழிபெயர்ப்புப் பணிக்குக் கூட ஒருவருக்கு இருமொழியிலும் திறமை இருக்கவேண்டும் என்ற அவசியம் இல்லை.

ஒரு மொழியின் வளம், தன்மை ஆகியவற்றைக் கணிப்பொறிக்கு ஏற்றாற்போல் கணிதப்படுத்துவது என்பது எளிமையான பணி அல்ல. எனவே, கணிப்பொறி வழி மொழி பெயர்ப்புத் திட்டங்கள் சுமார் 50 ஆண்டுகளுக்கு முன்னர் தொடங்கப்பட்டிருப்பினும் இன்னும் முழுமைபெறவில்லை. மேலும் கணிப்பொறி அறிவியலில் தொழில் நுட்ப வளர்ச்சி கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிக்கு மகிவும் பயனுள்ளதாக அமைகின்றது. இத்தகைய பணிகளின் நிறை குறைகளைக் கீழேக் காணலாம்.

மொழிக்கு இடைடே உள்ள சொல் வளம் இலக்கணம் கருத்தாடல் போன்றவைகளின் ஒற்றுமைகளை ஆராய வேண்டும். இவைகளில் மொழிக்கு இடையே உள்ள ஒற்றுமை அதிகரிக்க அதிகரிக்க மொழிபெயர்ப்புப் பணியின் சிக்கல்கள் குறையும்.

உருப்பனியல் அமைப்பை பொறுத்தவரையில் மொழிகளின் பண்புகள் இருவேறு திசைகளில் மாறுகின்றன:

அ. ஒரு சொல்லில் எத்தனை உருபன்கள் உள்ளன என்பதைப் பொறுத்து.

ஆ. ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இருந்தாலும் அதற்கு மேலே இருந்தாலும் அவைகளைத் துல்லியமாகப் பிரிக்க இயலும் என்பதைப் பொறுத்து.

ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இந்தால் அந்த மொழியைத் தனிநிலை உருபன்கட்டு மொழி (isolating Language) மொழி என்கிறோம். ஒரு மொழியில் உள்ள சொற்களில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருந்தால் அதை பன்நிலை உருபன் கட்டுமொழி (Poly Synthetic) என்கிறோம். சிபேரியன் மொழி பன்நிலை உருபன் கட்டுமொழிக்குச் சிறந்த உதாரணமாகும். வியட்நாமிஸ் மற்றும் கேண்டுனஸ் ஆகிய மொழிகள் தனிநிலை உருபன்கட்டு

(isolating) மொழிக்குச் சிறந்த உதாணரங்களகும். எனவே இத்தகைய மொழிக்குச் மொழி மாற்றம் செய்ய கணிப்பொறி வழி திட்டங்கள் வகுக்கின்ற பொழுது மொழியின் உருபனியல் பற்றிய விரிவான தகவல் அவசியம் ஆகும்.

மொழியின் தொடரியல் அமைப்புகள் (Syntactic Structure) மூன்று வகைகளாய் பிரிக்கப்படுகின்றன.

1. SVO அமைப்பு

ஜெர்மன், பிரஞ்சு, ஆங்கிலம் போன்றவை SVO அமைப்பில் உள்ள மொழிகளாகும்.

2. SOV அமைப்பு

தமிழ், இந்தி, ஜப்பான் மொழிகள் SOV அமைப்பில் உள்ள மொழிகளாகும்.

3. VSO அமைப்பு

இர்ரீஸ், அரேபி, ஷிபிரோ போன்ற மொழிகள் VSO அமைப்பில் அமைந்துள்ளன.

SVO அமைப்பில் உள்ள மொழிகளில் முன்னுருப்புகள் (Preposition) பயன்படுத்தப்படுகிறது. எனவே கணிப்பொறி வழி மொழிபெயர்ப்புத் திட்டங்களுக்கு மொழிகளுக்கு இடையே உள்ள தொடரியல் அமைப்பு வகைகள், உறவுகள் போன்றவை ஆராயப்பட வேண்டும்.

இராமனின் வீடு என்ற தொடரில் *வீடு* என்பது தலைச்சொல் (Head Word) ஆகும். *இராமன்* என்பது சார்புச் சொல் (Dependent Word) ஆகும். தலைச்சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையில் உள்ள உறவை -*இன்* என்ற ஒட்டு விளக்குகிறது. இது போன்று தலைச் சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையிலான முறையைப் பொறுத்து மொழி தலைக்குறித்தல் மொழி, சார்புக்குறிதல் மொழி எனப் பிரிக்கப்படும். சார்பு குறித்தல் மொழியில் ஒட்டு சார்புச் சொல்லுடனும் தலைக்குறித்தல் மொழியில் ஒட்டு தலைச் செல்லுடனும் தரப்படுகின்றன. ஆங்கில மொழி தலைக்குறித்தல் மொழியைச் சார்ந்ததாகும். ஹங்கேரிய மொழி சார்புக்குறித்தல் மொழியைச் சார்ந்ததாகும். பின்வரும் எடுத்துக்காட்டு விளக்கும்.

English	:	The Man’s house
Hungarion	:	The man house his
House	:	Head word (தலைச்செல்)
The Man	:	Dependent word (சார்புச்சொல்)
-s	:	Affix (ஒட்டு)

எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியை மேற்கொள்கின்ற போது தலைச்சொல்லுக்கும், சார்பு சொல்லுக்கும் இடையே உறவு முறைகள் எவ்வாறு குறிக்கப் பெறுகின்றன என்பது பற்றியெல்லாம் ஆராயவேண்டும்.

மூலமொழியில் உள்ள ஒரு சொல்லின் வகைப்பாடும் இலக்கு மொழியிலுள்ள அதற்கு நிகரான சொல்லின் வகைப்பாடும் ஒன்றாக இருக்கவேண்டியத் தேவையில்லை. எடுத்துக்காட்டாக ஒரு மொழியிலுள்ள வினைச்சொல் பிறமொழிக்கு மொழி பெயர்ப்பு செய்யப்படும் போது மொழிப்பெயர்ப்புச் சொல் வினைச்சொல்லாக இருக்கவேண்டிய தேவையில்லை. இவைகள் எல்லாம் மொழிபெயர்ப்புப் பணியில் உள்ள சிக்கல்களாகும்.

ஒவ்வொரு மொழியிலும் கருத்தாடல்களுக்கானச் சொல் விளக்கமானதாகவோ கருக்கமானதாகவோ இருக்கலாம். எடுத்துக்கட்டாக *சகோதரன்* என்ற உறவைக் குறித்து ஆங்கிலத்தில் *brother* என்ற சொல் உள்ளது. மூத்த சகோதரனை ஆங்கிலத்தில் *elder brother* என்று சொல்லுகிறோம். இவ்வாறு வயது மூத்தவர் இளையவர் என்பதைக் குறிக்க *elder*, *younger* என்ற அடைச் சொற்கள் பயன்படுத்தப்படுகின்றன. ஆனால் தமிழில் *அண்ணன்* என்றும் *தம்பி* என்றும் தனித்தனிச்சொற்கள் உள்ளன. *சாகோதரன்* என்ற உறவை மேலும் வகைப்படுத்த முதியவர், இளையவர் என்ற நிலையில் *அண்ணன்*, *தம்பி* என்ற இருச்சொற்கள் உள்ளன. ஆனால் ஆங்கிலத்தில் இவ்வகையிலான வேறுபாடு என்பது இல்லாதால் உறவை வேறுப்படுத்த *elder*, *younger* என்ற சொற்களைப் பயன்படுத்த நேர்கிறது.

மூலமொழியில் உள்ள செய்திகளில் கலாசாரப் பின்னணி மொழிப்பெயர்ப்பிற்கு ஒரு முக்கியமான காரணியாகும். அதே கலாச்சாரப் பண்பு இலக்கு மொழியின் கலாச்சாரப் பண்புகளுக்கு ஏற்ப மாற்றித்தரத்தக்க அளவு இருத்தல்வேண்டும். இருமொழிகளுக்கு இடையே உள்ள கலாச்சார தொடர்புகள் மொழிபெயர்ப்பு செய்ய இயலுமா? இயலாதா? என்பதைக் தெளிவுபடுத்த உதவும். மேலும் சபீர்-ஊர்ஃப் கொள்கையின் படி ஒருவர் பேசும் மொழி சிந்திக்கும் திறனைக் கூட்டுவதாகவோ குறைப்பதாகவோ உள்ளது. எனவே மொழிபெயர்ப்பில் இத்தகைய உளவியல் அடிப்படையிலானச் சிக்கல்கள் உள்ளன.

தொடரியல் பகுப்பாய்வி பல பணிகளுக்காக உருவாக்கப்படுகின்றன. ஆனால் இயந்திர மொழிபெயர்ப்பிற்காகப் பயன்படுத்தப்படும் பகுப்பாய்வு மற்ற பகுப்பாய்வை விட சிறிது வேறுபடுகிறது. இங்குத் தொடரின் அல்லது வாக்கியத்தின் பொருளை அடிப்படையாகக் கொண்டு பகுப்பாய்வு செய்யப்படுகிறது. எனவே இயந்திர மொழிபெயர்ப்பிற்கு உருவாக்கப்படும் பகுப்பாய்வு தொடரியல் அமைப்பை மட்டும் சாராது பொருளைக் சார்ந்ததாகவும் அமைகிறது. இத்தகைய பகுப்பாய்வுகளை உருவாக்க அறிவு அடிப்படை (Knowledge Base) தேவைப்படுகின்றது.

இயந்திரம் வழி மொழிபெயர்ப்பு செய்கின்ற பொழுது ஏதேனும் சிக்கல்கள் இருந்தாலோ அல்லது கணிப்பொறிக்குக் கூடுதல் கதவல் தேவைப்பட்டாலோ மனித உதவியை எதிர்நோக்குகின்றது. இத்தகைய சூழல்களில் சிக்கலைப் புரிந்துக்கொண்டு அதற்கேற்றத் தீர்வு உரிய முறையில் வழங்கப்பட்டவேண்டும். எனவே இயந்திர மொழிப்பெயர்ப்புத் திட்டத்தை இயந்திர மனித உரையாடலுக்கு ஏற்றாற்போல் வடிவமைக்கவேண்டியது அவசியம் ஆகும். பெரும்பாலும் படைப்பாற்றல் திறனைச் செய்திகளில் வெளிப்ப்படுத்தும் பொழுது கணிப்பொறிக்குக் கூடுதல் செய்திகள் தேவைப்படுகிறன்றன.

பெரும்பாலான கணிப்பொறி வழி மொழிபெயர்ப்புகளிலும் உள்ளீடு செய்யும் செய்திகளை முன்திருத்தம் செய்யும் பொழுது மூலமொழியின் தன்மையும் இலக்கு மொழியின் தன்மையும் கருத்தில் கொள்ளப்படுகின்றன. மூலமொழியில் உள்ள செய்தியை முதலில் இலக்கு மொழிக்கு மொழிபெயர்ப்பது, பின்னர் மொழி பெயர்ப்பைப் படித்தறிந்து அது சரியான கருத்தை வெளிப்படுத்துகின்றதா என்பதை ஆராய்வது, அப்படி இல்லையெனில் மூலமொழி செய்தியை முன்திருத்தம் செய்து மறுபடியும் மொழிபெயர்ப்பு செய்வது, சரியான கருத்து வெளிப்படுவதுவரை இத்தகைய ஆய்வை மீண்டும் மேற்கொண்டு மொழி பெயர்ப்பிற்கு முன்னால் முன்திருத்தப் பணிக்கு முக்கியத்துவம் அளித்து தேவையான திருத்தங்களைச் செம்மையாக மேற்கொள்வது. இம்முறை கணிப்பொறியின் நேரத்தை மிச்சப்படுத்துகிறது; மொழிபெயர்ப்பின் வேகத்தை அதிகப்படுத்துகிறது. மேலும் ஒரு மொழியிலிருந்து பல மொழிகளுக்கு மொழிபெயர்ப்பு செய்ய வேண்டிய சூழல்களில் முன்திருத்தப்பணித் திறமையானதாக அமைகிறது.

மொழிபெயர்ப்புப் பணிகளுக்காக உருவாக்கப்படுகின்ற அகராதிகள் சாதாரண முறையில் அல்லாது பல கூடுதல் செய்திகளைக் கொண்டதாக இருத்தல் வேண்டும். பிறவகை அகராதிகளை விட துறை சார்ந்த அகராதிகள் மொழிபெயர்ப்புப் பணியை எளிமைப்படுத்துகிறது. மூலமொழியில் உள்ள செய்திகளைப் பொறுத்து அது தொடர்புடைய அகராதியைப் பயன்படுத்திக்கொள்ளலாம்.

மொழிபெயர்ப்புக்காக உருவாக்கப்படும் அகராதியில் சொற்கள், தொடர்கள், மரபுத் தொடர்கள் மேலும் அடிக்கடி பயன்படுத்தக்கூடிய வாக்கியங்கள் போன்றவைகள் உரிய விளக்கத்துடன் தரப்படவேண்டும்.

கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணி என்பது ஒரே மூச்சில் ஒட்டு மொத்தமாகச் செய்யக்கூடிய பணியாகும். முதலில் கணிப்பொறி வழியமைப்பிற்கான மென்பொருள் வடிவமைக்கப்படவேண்டும். அதன் செயல் திறனை அறியவேண்டும். பணியின் தன்மையை மதிப்பிடவேண்டும். இதற்கு ஏற்றார் போல் தேவைப்படின் மென்பொருளைக் கூடுதல்

வசதிகளுடன் மேம்படுத்தப்பட வேண்டும். இம்முறைகளை மனிதனுக்கு நிறைவழிக்கும் கொள்திறன் கிடைக்கின்ற வரையில் திரும்பத் திரும்பச் செய்ய வேண்டும்.

கணிப்பொறி வழி மொழிபெயர்ப்புக் கடுமையானப் பணியாகும். இதற்கு மனிதமொழி பெயர்ப்பில் மேற்கொள்ளப்படும் திட்ட மாதிரிகளும் (Model) வழிமுறைகளும் அவசியமாகும். கணிப்பொறி வழி மொழிபெயர்ப்பு எண்ணற்ற முயற்சிகளால் உருவாக்கப்படுகின்ற ஆக்கப்பூர்வமான பணியாகும். இதற்கு அறிவியல் திறன் மட்டுமல்லாது மொழிபெயர்ப்பிற்குத் தேவையான கலைத்திறனும் அவசியமாகும்.

தற்பொழுது உலகளவில் நடைபெற்றுள்ள மொழிபெயர்ப்புப் பணிகள் மனதுக்கு நிறைவழிக்கக்கூடிய வகையில் உள்ளன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்கு உதவுகின்றன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்குப் போதுமானதாகும்.

மனிதனின் படைப்பாற்றல் கற்பனைத்திறன், உளவியல்பாங்கு போன்ற செய்திகளால் பொதிந்துள்ளதால் கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியில் இத்தகையக் கூறுகளைச் சேர்க்கவேண்டி உள்ளது. அதற்கு மனித நுண்ணிறிவுடன் கூடிய கணிப்பொறிகள் உருவாக்கப்படுகின்றன.

10.7. இயந்திர மொழிபெயர்ப்பின் வளர்ச்சி

இயந்திர மொழிபெயர்ப்பின் வளர்ச்சியை நான்கு காலகட்டமாகப் பிரித்து பகுத்து ஆயலாம் (Hutchins ,1986). அவை முதல் காலக்கட்டம், இரண்டாவது காலகட்டம், மூன்றாவது காலகட்டம், நான்காவது காலகட்டம், இன்றைய காலகட்டம் எனக் கூறப்படும்.

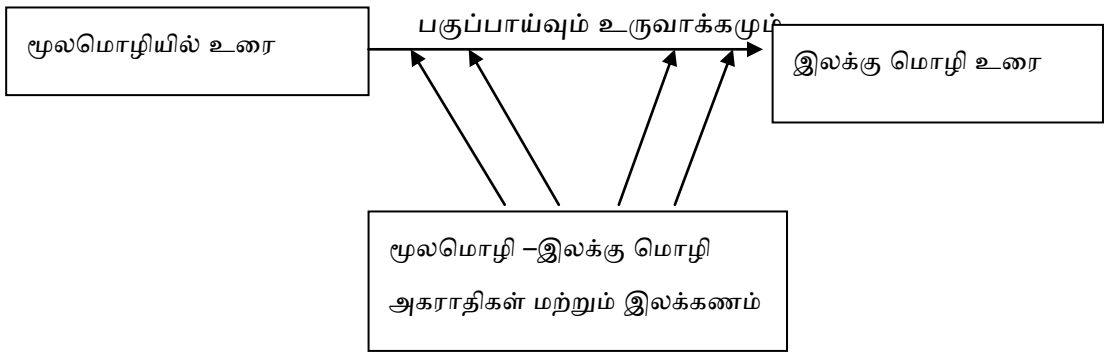
முதல் கட்டம் (1946லிருந்து 1982 வரை)

இக்காலகட்டத்தில் செய்யப்பட்ட இயந்திர மொழிபெயர்ப்புகள் மூல மொழிச் சொற்களுக்கு இணையான சொற்களை இருமொழி அகராதி பார்த்துக் கண்டுபிடித்து மூல மொழி போலவே வரிசைப்படுத்தித் தருவதைச் செயல்முறையாகக் கொண்டிருந்தது. இச்செயல்முறை தொடரியல் பகுப்பானை பயன்படுத்தவில்லை.

இரண்டாவது காலகட்டம் (1954-1966)

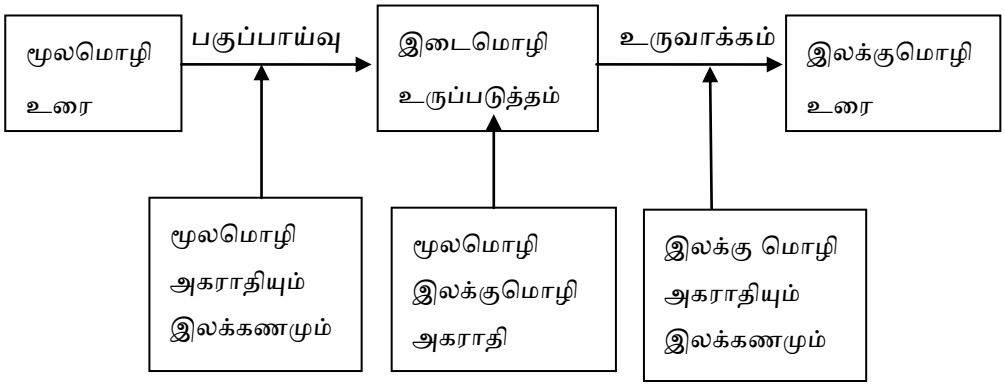
1954-இல் ஜார்ஜ்டவுன் பல்கலைக்கழகம் 880 ரஷ்ய சொற்களையும் இலக்கணத்தின் ஆறு விதிகளையும் கொண்டு ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை நடைமுறைப்படுத்திக் காட்டியது. அடிப்படைச் சிக்கல் பொறியியல் தன்மை அடிப்படையிலான சிக்கல் என்று நம்பப்பட்டது. அடுத்த பத்தாண்டுகளில் யு.எஸ்.ஏயில் (USA) இத்தகைய ஆய்வுகள் பெருமளவில் ஊக்குவிக்கப்பட்டன. கடினமான மொழியியல் சிக்கல்கள் குறித்த விழிப்புணர்வு அதிகரித்தது. திருப்தியில்லாத முன்னேற்றம் காரணமாக ALTPA இயந்திர மொழிபெயர்ப்பிற்கான பண உதவியைக் குறைக்கப் பரிந்துரைத்தது. இக்காலகட்டத்தில் தொடரியல் ஆய்வு தொடக்க

நிலையில் இருந்தது. இலக்கண விதிகள் (grammatical Rules) மற்றும் தொடரியல் அமைப்பு (syntactic structure) ஆகிய கருத்துச்சாயல்கள் இல்லாதிருந்தது. ஆங்கிலம் மற்றும் ருஷ்ய இலக்கணங்கள் வழியமைப்பின் (program) ஒரு பகுதியாக இருந்தன.



மூன்றாவது காலகட்டம் (1966-1975)

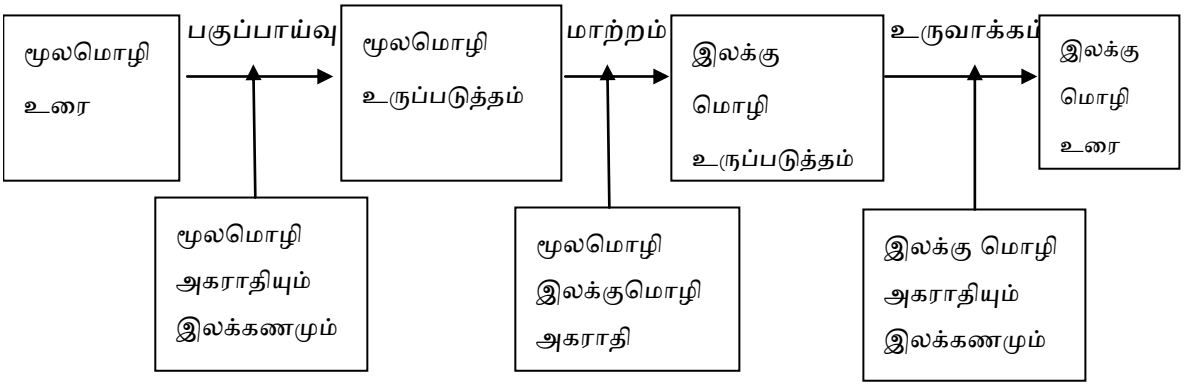
நேரடியான மொழிபெயர்ப்பு அடிப்படையில் இயங்கும் இரு ஒழுங்கமைப்புகள் Systran மற்றும் Logos. இக்காலகட்டத்தில் மொழியியல் கோட்பாட்டின் முன்னேற்றத்தில் மிகுந்த மாற்றம் ஏற்பட்டது. மொழியியல் தரவு கணிப்பொறியின் செயல்பாட்டிலிருந்து பிரிக்கப்பட்டது. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் இடைமொழி அணுகுமுறையைப் (Interlingual Approach) பயன்படுத்தியது; இதன் செயல்பாட்டில் இரண்டு நிலைகள் இருந்தன: மூல மொழியிலிருந்து இடைமொழிக்கு மற்றும் இடைமொழியிலிருந்து இலக்கு மொழிக்கு.



Metal மற்றும் Ceta என்ற இரண்டு ஒழுங்குமுறைகள் இடைமொழி அணுகுமுறையின் (Interlingual Approach) அடிப்படையில் உருவாக்கப்பட்டவை. இவ்வணுகுமுறையில் மூல

மொழிக்கும் இலக்கு மொழிக்கும் அவைகளுடைய தனித்தன்மையான அக அமைப்புகளின் உருப்படுத்தம் (Deep Structure Representation) இருக்கும். மொழிபெயர்ப்பு மூன்று மட்டச் செயல்முறையாகும்.

1. உரைகளை மூல மொழியின் உருப்படுத்தமாகச் செய்தல்
2. அதை இலக்கு மொழி உருப்படுத்தமாக மாற்றுதல்
3. இலக்கு மொழியின் உரைகளை உருவாக்குதல்



மாற்ற ஒழுங்குமுறையின் மிக அறியப்பட்ட எடுத்துக்காட்டு மாண்ட்ரிஸ் பல்கலைக்கழகத்தில் செய்யப்பட்ட ஆங்கில பிரெஞ்சு மொழி ஒழுங்குமுறையான TAUM திட்டம் ஆகும்.

நான்காவது காலகட்டம் (1975-1985)

இந்த காலகட்டத்தில் உருவாக்கப்பட்ட ஒழுங்குமுறை குவிதல் அணுகுமுறைகளைப் (convergence approach) பயன்படுத்தியது. இவைகள் கட்டுப்படுத்தப்படாத உரை உள்ளீடு கொண்ட முழு தானியங்கு ஒழுங்குமுறைகளின் திட்ட அமைப்பிற்கு எல்லைப்படுத்தப்பட்டன. இந்தக் காலகட்டத்தில் எல்லைக்குட்பட்ட மொழி ஒழுங்குமுறைகளும் ஊடாடும் ஒழுங்குமுறைகளும் உருவாக்கப்பட்டன. தொடரியல் மற்றும் பொருண்மையியலுக்கு எல்லைப்படுத்தப்பட்ட ஒழுங்குமுறையின் எடுத்துக்காட்டு TAUM என்பதன் அனுபவத்தின் அடிப்படையில் மாண்டரியஸ் பல்கலைக்கழகத்தால் உருவாக்கப்பட்ட Meteo என்பதாகும். இது ஆங்கிலத்தில் உள்ள வானிலை அறிக்கைகளைப் பிரெஞ்சு மொழிக்கு மொழிபெயர்க்கும். ஊடாட்டம் செய்யும் இயந்திர மொழி பெயர்ப்பின் கவர்ச்சி மனித மொழிபெயர்ப்பாளர்கள் மற்றும் கணினி இவற்றின் ஒத்துழைப்பைச் சிறப்பாகப் பயன்படுத்துவதில் உள்ளது.

இன்றைய கால கட்டம் (1985-இல் இருந்து)

ஆளுகைக் கட்டுறவுக் கோட்பாடு (Government and Binding (GB) Theory) , சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG)), கிளை இணைக்கும் இலக்கணம்

(Tree Adjoining Grammar (TAG)), பெரிதாக்கப்பட்ட மாற்ற வலைப்பின்னல் (Augmented Transition Network (ATN)) போன்ற புதிய இலக்கண வடிவங்களின் வருகை இயற்கை மொழி ஆய்வுப் பரப்பில் புதிய பார்வையை உருவாக்கியது. இம்மொழியியல் வடிவமைப்புகள் பகுத்துக்குறிக்கும் வழிமுறைகளில் தொடரியல் மற்றும் பொருண்மையியல் பண்புக்கூறுகளின் பொருத்தத்திற்கு ஆலோசனைக்குரியது. மேலும் கணிப்பொறி அறிவியலில் ஏற்பட்ட தொழில்நுட்ப முன்னேற்றம் இயந்திர மொழிபெயர்ப்பு ஆய்வின் வளர்ச்சிக்குச் சாதமாக அமைந்தது. சொல்-பொருண்மையியல் பகுப்பாய்வை உள்ளடக்கிய இயற்கை மொழியின் புரிதல் (Natural Language Understanding) தேவையான உலக அறிவை அடையாளம் காணுவதை உள்ளடக்கியது. அறிவு உருப்படுத்தம் மற்றும் மொழியியல் அறிவில் உள்ளடக்குவது போன்றவை இயந்திர மொழிபெயர்ப்பு ஆய்வுகளின் பரப்பில் மேலும் கூடுதல் வளர்ச்சிக்கு உதவி புரிந்தது. தற்கால ஒழுங்குமுறைகள் இந்த மூன்று அணுகுமுறைகளின் மேம்பாடுகளைப் பயன்படுத்துகின்றன. முழு ஒழுங்குமுறையும் பல எண்ணிக்கையிலான துணை ஒழுங்குமுறைகளாகப் பிரிக்கப்பட்டுச் சரியாக இடைமுகப்படுத்தப்பட்டுள்ளன. இயந்திர மொழிபெயர்ப்பின் இவ்வேறுபட்ட கட்டங்களை மூன்று நிலைகளாக் கூறலாம்.

1. **ஆய்வு நிலை**

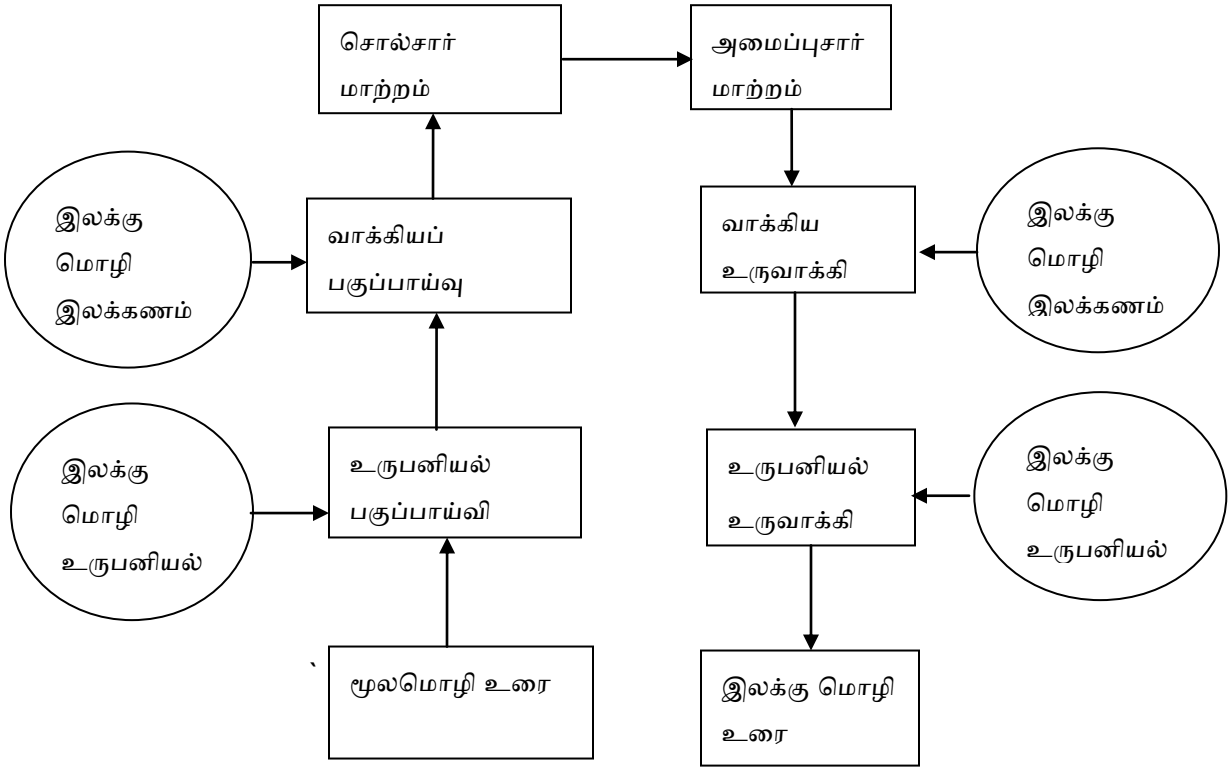
ஒழுங்குமுறையின் உள்ளீடாக மூல மொழி உரை முதலில் சொல் நிலையில் ஆயப்படுகிறது. உருபனியல் ஆய்வு, உருபனியல் விதிகள், சந்தி விதிகள், அடிச்சொல், இயந்திரம் படிக்கவியலும் அகராதி மற்றும் முன்னொட்டு அகராதி போன்ற பலதரப்பட்ட அகராதிகளைப் பயன்படுத்திச் சொற்களை அவற்றின் உருபன்களுக்காகப் பிரிக்கின்றது. வாக்கியப் பகுப்பாய்வியில் ஒரு இலக்கண மாதிரியைப் பயன்படுத்தி ஒவ்வொரு வாக்கியத்தையும் அவற்றின் உறுப்புக்கான தொடர்கள் மற்றும் எச்சத் தொடர்களுக்காகப் பகுப்பாய்வு செய்கிறது; பயனிலைகளுக்கும் அவற்றின் பங்கெடுப்பாளர்களுக்கும் இடையில் உள்ள தொடரியல் மற்றும் பொருண்மையியல் உறவுகள் குறியாக்கம் செய்யப்படுகின்றன.

2. **மாற்ற நிலை**

இந்நிலையில் இரண்டு நிலைகள் உள்ளன: 1. சொல்சார் மாற்றம் 2. அமைப்புசார் மாற்றம். சொல்சார் மாற்ற நிலையில் உரை மற்றும் அகராதியின் தொடரியல் மற்றும் பொருண்மையியல் குறியாக்கத்தின் உதவியுடன் ஒவ்வொரு சொல்லின் சொல்சார் பொருண்மையின் மெய்ப்படுத்தம் செய்யப்படும்; இலக்கு மொழியிலிருந்து பொருத்தமான சொல் தேர்ந்தெடுக்கப்படும். அமைப்பு மாற்ற நிலையில் இலக்கு மொழி, மூல மொழி விதிகளைப் பயன்படுத்தி மூல மொழி அமைப்பு இலக்கு மொழி அமைப்பிற்கு மாற்றப்படும்.

3. **உருவாக்க நிலை**

இந்த நிலையில் வாக்கிய உருவாக்கும் பகுதி (Sentence Generator Module) இலக்கு மொழி இலக்கணத்தின் உதவியால் பெறப்படும் அமைப்புகளுக்கு இலக்கண அடிப்படையில் சரியான வாக்கியங்களை உருவாக்கும். பின்னர் உருபனியல் விதிகள் அல்லது புணர்ச்சி விதிகள், இயந்திரம் படிக்கவியலும் முன்னொட்டு அகராதி இவற்றைப் பயன்படுத்தி உருபனியல் சார் குறியாக்கம் இலக்கு மொழியின் பொருத்தமான ஒட்டுகளால் இடம் பெயர்க்கப்படும். வெளியீடு இலக்கு மொழியில் ஒரு உரையாக உற்பத்தி செய்யப்படும். இம்மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் வேறுபட்ட நிலைகளும் அவற்றின் துணை ஒழுங்குமுறைகளும் கீழே தரப்பட்டுள்ளன.



10.7.1. இந்தியாவில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சி

ஆரம்பக்கட்டத்தில் இந்தியாவில் இயந்திர மொழிபெயர்ப்புகள் இதற்கான சாத்தியத்தை வெளிப்படுத்தும் பொம்மை ஒழுங்குமுறைகளாக (toy systems) இருந்தன. 1978 வரை இயந்திர மொழிபெயர்ப்பிற்கான எந்த முயற்சியும் எடுக்கப்படவில்லை. 1978-இல் ஐ.ஐ.டி.கான்பூரில் உள்ள மின்னியல் தொழில் நுட்பத் துறையின் முயற்சியால் மொழியியல் அடிப்படையாலன தகவல் ஒழுங்குமுறைகள் பற்றிய தேசியக் கருத்தரங்கு (National Symbosium on Linguistic Based Information System) முதன் முதலில் நடைபெற்றது. இக்கருத்தரங்கு

ஆராய்ச்சியாளர்களிடையேயும் அரசாங்கத்திலும் தொழில்நுட்ப துறைகளிலும் இயந்திர மொழிபெயர்ப்பு குறித்த விளிப்புணர்ச்சியை ஏற்படுத்தியது. இருப்பினும் இயந்திர மொழிபெயர்ப்பு 80-களின் ஆரம்பகட்டத்தில் இந்திய மொழிகளுக்குச் செல்லாய்விகளும் பிற கணிபொறி வசதிகளும் வரத் தொடங்கிய பின்னர்தான் தொடங்கப்பட்டது. ஐ.ஐ.டி.கான்பூர் GIST என்ற தொழில் நுட்பத்தின் உருவாக்கத்திற்கு முயன்றது. இது இயந்திர மொழிபெயர்பிற்கு விழிப்புணர்ச்சி ஏற்படுத்துவதுடன். இதில் ஆய்வு மற்றும் வளர்ச்சி குறித்த பல திட்டங்களை ஊக்குவித்தது. இந்தியாவில் பல மையங்களில் பலவிதமான முயற்சிகள் மேற்கொள்ளப்பட்டாலும் மூன்றும் அணுகுமுறைகளைப் பட்டியலிட இயலும்.

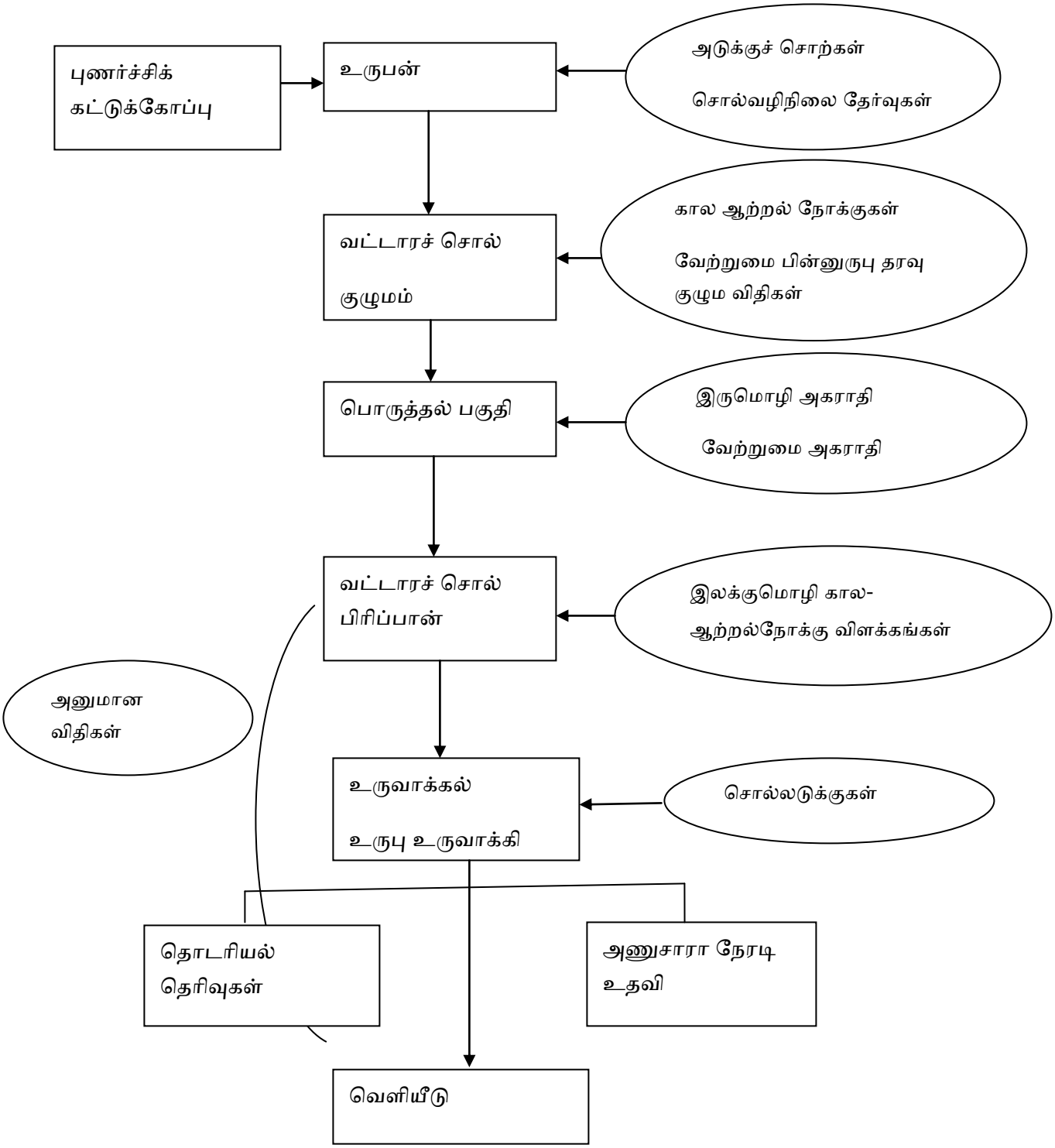
இடமொழி அணுகுமுறை (Interlingual Approach)

நேரடி சொல்சார் மாற்றம் (Direct lexical Transfer Approach)

கணிப்பொறி உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிகள் (Machine Aided Translation Effort)

10.7.1.1.அனுசாரக் இயந்திர உதவியிலான மொழிபெயர்ப்பு ஒழுங்குமுறை

சமஸ்கிருதத்தை இடைப்பட்ட அடிப்படை மொழியாகக் கொண்டு இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்ப்பு செய்யும் இடைமொழி ஆய்வு அணுகுமுறை சின்ஹா என்பவரால் 1984-இல் முயற்சிக்கப்பட்டு 1989-இல் விரிவாக்கப்பட்டது. இது மூல மொழியின் ஆய்வுக்குக் காரக அடிப்படையிலான அக உருப்படுத்தத்தை முன்மொழிந்தது. 1986-1988-களில் எல்லைகுட்பட்ட சொற்களைக் கொண்ட மிக எளிய வாக்கியங்களை இந்தியிலிருந்து தெலுங்கிற்கு மொழிபெயர்க்கும் மூலமுன்மாதிரி (prototype) சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் உருவாக்கப்பட்டது. இது காரக அடிப்படையிலான கருத்துவரைபடத்தை அக உருப்படுத்த்திற்கும் பொருண்மை மயக்கங்களுக்கும் பயன்படுத்தியது. ஒரு தனிநிலை தொடரியல் இருப்பானில் உள்ள மிகக்கூடுதலாகத் தொடர்புள்ள சொற்கள் ஒரு சொல் குழுவாகப் பயன்படுத்தப்பட்டது. மூல மொழியில் உள்ள உரை உருவாக்கி இச்சொற் குழுமங்களிலிருந்து பொருத்தத்தைச் சரியான உருபனியல் உருவாக்கத்திற்குப் பயன்படுத்தியது.



இடைமொழி அடிப்படையிலான இயந்திர மொழிபெயரிப்பு அதற்குப்பின் முயற்சிக்கப்படவில்லை. இவ்வணுகுமுறை இந்திய மொழிகளின் பன்மைத்தன்மைக்குக் கூடுதல் அனுகூலமாய் இருந்தாலும் அக உருப்படுத்ததிலிருந்து புற அமைப்பை பெறுவது கடினமான

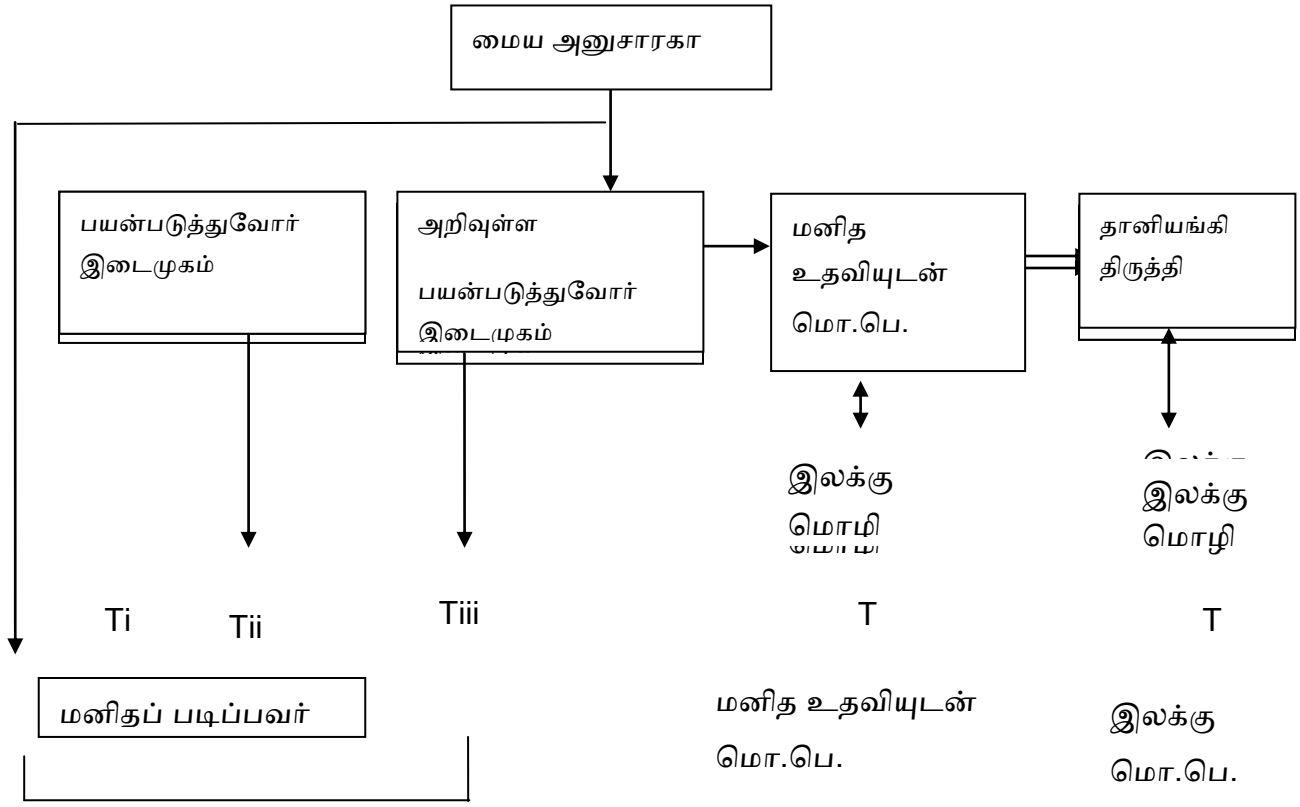
செயல்பாடாக அமைந்து குறையுடையதாய் இருந்தது. இவ்வணுகுமுறை மூலமொழியின் சீரான புரிதலை முன்கருதலாகக் கொண்டதால் இன்றைய ஆய்வுநிலையில் ஏற்றதாக அமையவில்லை. மேலும் அமைப்பு அடிப்படையில் அண்மைப்படுகின்ற இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்பின் போது நேரிடையான சொல் பதிலீடு செய்வது எளிமையாக அமைந்தது. இதன் காரணமாக இவர்களால் அனுசாரகா என்ற அணுகுமுறை உருவாக்கப்பட்டது.

இந்திய மொழிகளுக்கிடையே உள்ள அமைப்பு ஒற்றுமை காரணமாக மூலமொழியிலிருந்து இலக்கு மொழிக்குப் பல பொருண்மை மயக்கங்களைக் கொண்டு செல்ல இயலும் என்ற காரணத்தால் இந்த இடைமொழி அணுகுமுறை இந்திய மொழிகளுக்கிடையே மொழிபெயர்ப்பிற்கு நல்லதொரு அணுகுமுறையாக அமையும். இருப்பினும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கும் இந்திய மொழிகளிலிருந்து ஆங்கில மொழிகளுக்கும் மொழிபெயர்க்க இடைமொழி அணுகுமுறை பொருத்தமான அணுகுமுறை அல்ல.

இயந்திர உதவியிலான மொழிபெயர்ப்பு ஒழுங்குமுறையான அனுசாரக (ANUSAARAKA) சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் மூலமொழியிலிருந்து இலக்கு மொழிக்கு நேரடி சொல்சார் மாற்ற அணுகுமுறையாக உருவாக்கப்பட்டது. இங்கு மூல மொழிக்கும் இலக்கு மொழிக்கும் இடையில் உள்ள வாக்கிய அமைப்பின் பொதுமைப் பண்புகள், இரண்டும் இந்திய மொழிகள் என்ற காரணத்தால் கருதப்பட்டு முழு அளவிலும் பயன்படுத்தப்பட்டுள்ளது. இவ்வணுகுமுறையின் படி ஒரு தனிநிலை தொடரியல் உறுப்பாக நடத்தை செய்யும் மூலமொழியிலுள்ள சொற்கள் குழுமப்பட்டு மூலமொழிக்குப் பொருந்தும்படி அது சொற்களால் பதிலீடு செய்யப்படும் பின்னுருபை நிர்ணயிக்க சில் விதிகள் உருவாக்கப்பட்டன. சொற்களின் நிரல் வருகை பெரும்பாலும் தக்க வைக்கப்பட்டு சில சமயங்களில் புதிய இணைப்பான்களின் பயன்பாட்டால் நிறைவு செய்யப்படும். சில சமயங்களில் இம்முறை வெளியீட்டில் தொடரியல் தவறுகளுக்கு வழிவக்கும்; இது பற்றி இதை உருவாக்கியவர்கள் கவலைப்படவில்லை. இருப்பினும் குறைந்த அளவு பொருள் பெரும்பாலும் வெளிப்படுத்தப்படும். மேலும் அனுசாரக்காவின் பாகங்கள் அல்லது உறுப்புத் தொகுதிகள் எந்த இயந்திர மொழிபெயர்ப்பிற்கும் அல்லது இயந்திர உதவியுடன் செய்யப்படும் மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கும் பகுதியாக அமையும்.

இவ்வாறு அனுசாரகாவின் வெளியீடு இடைப்பட்ட ஒரு நிலைக்கு முக்கிய வெளியீடாகும் மற்றும் அனுசாரகா போன்ற வெளியீடு ஒரு இரண்டாம் நிலை விளைவாகும் (by product). இது சீரற்ற மொழிபெயர்ப்பிற்கு கொண்டுசெல்லாததால் மொழிகளின் இணையின் இயந்திர மொழிபெயர்ப்பிற்கும் பயனுள்ள உபாயமாகும். இதை உருவாக்கியவர்கள் இதை ஒரு மொழி அணுகி(language acessor) என்று அழைகின்றனர்; இதிலிருந்து முழு நன்மை அடைய பயன்படுத்துபவர்களிடம் சில பயிற்சியை எதிர்பார்க்கின்றனர். தொடக்கத்தில் கன்னடா, இந்தி

மொழிகளுக்கு ஒரு அனுசாரக் ஒழுங்குமுறை அமைக்கப்பட்டு காட்டப்படது. இதன் பின்னர் தெலுங்கு-இந்தி, பஞ்சாபி-இந்தி, மராட்டி-இந்தி, சமஸ்கிருதம்-இந்தி, தமிழ்-இந்தி ஆகிய மொழி இணைகளுக்கு அனுசாரக் ஒழுங்குமுறைகள் உருவாக்கப்பட்டு பார்வையாளர்களுக்கு எடுத்துக்காட்டப்பட்டது.



அனுசாரகா வெளியீட்டில் பல மட்டங்கள்

(http:/iiit.net/ltrc/Anusaaraka/anu_home.html).

10.7.1.2.சிவ மற்றும் சக்தி ஒழுங்குமுறை (SHIVA and SHAKTI MT System)

யு.எஸ்.எ.இல் உள்ள கார்னெகி மெல்லொன் பல்கலைகழகமும் பெங்களூரில் உள்ள இந்திய அறிவியல் நிறுவனமும் ஹெதராபாத்திலுள்ள உலகத் தகவல் தொழில்நுட்ப நிறுவனமும் இணைந்து ஆங்கிலத்திலிருந்து இந்திக்கு மொழிபெயர்ப்பதற்காக சிவ (SHIVA) மற்றும் சக்தி (SHAKTI) என்ற இரண்டு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்கியது. இவ்விரு மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் பரிசோதனைக்காகவும் வெள்ளோட்டமாகவும் பயன்படுத்துபவரின் கருத்தறியவும் வெளியிடப்பட்டன. சிவ என்ற மொழிபெயர்பு ஒழுங்குமுறை எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையாகும். (சிவ

ஒழுங்குமுறையை பின்வரும் வலை தளத்தில் பெறலாம்: (http://ebmt.serc.iisc.ernet.in/mt/login.html மற்றும் சக்தி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை பின்வரும் வலை தளத்தில் பெறலாம்: http://shakti.iiit.net). 2006 ஏப்ரலில் சக்தியின் 0.81 மாதிரி வெளியிடப்பட்டது. இது இந்தி, மராத்தி, தெலுங்கும் என்ற மூன்று இலக்கு மொழிகளுக்குச் செயல்புரியும். பயன்பாட்டாளர் ஆங்கில வாக்கியங்களை மொழிபெயர்க்க முதலில் இந்த இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையுடன் இணைய வேண்டும். ஒழுங்குமுறை பட்டியலிருந்து எழுத்துருவை தெரிந்தெடுக்கக் கேட்கும்; மொழிபெயர்க்கப்பட்ட வெளியீட்டை தெரிந்தெடுக்கப்பட்ட எழுத்துருவில் காட்டும். பயன்பாட்டாளர் மொழிபெயர்ப்பு ஒழுங்குமுறையால் உருவாக்கப்பட்ட மொழிபெயர்ப்பில் திருப்தி பெறவில்லை என்றால் ஒழுங்குமுறை பயன்பாட்டாளரிலிருந்து மூலமொழியின் பல்வேறு கூறுகளின் (சொல், தொடர், வாக்கிய நிலைகளில்) பொருள்களைக் கேட்டு உள்ளீட்டுப் பனுவலை மீள் மொழிபெயர்ப்பு செய்யும்.

10.7.1.3 ஆங்கிலபாரதி (*ANGLABHARATI*)

ஐ.ஐ.டி.கான்பூரில் மேற்கொள்ளப்பட்ட ஆங்கில பாரதி (*ANGLABHARATI*) திட்டத்தை நிறுவுதில் நான்கு முக்கிய கருதல்கள் இருந்தன. முதலாவது அன்றைய காலகட்டத்தின் நிலையில் சீரான இயந்திர மொழி பெயர்ப்பு சாத்தியமில்லை. எனவே இயந்திர உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிக்கப்பட வேண்டும்; இங்கு இயந்திரம் பெரும்பாலான செயல்களைச் செய்கின்றது. கடினமான வேலைககள் முந்தைய மற்றும் பிந்தைய திருத்தங்களால் (Pre and Post editings) கையாளப்படுகின்றது. இதன்படி ஒவ்வொரு மூலமொழி வாக்கியத்ததையும் தனியாகப் பரிசோதிப்பது மற்றும் முற்கட்டு மற்றும் (anaphora) வாக்கியங்களுக்கிடையிலான குறிப்புகள் (Intersentential References) போன்ற சிக்கல்களை மனிதத் திருத்துனருக்கு விட்டுவிடுவது என்பன எளிமைக்கு வழி வகுக்கிறது. இரண்டாவது இயந்திர மொழி பெயர்ப்பு, பயனப்படுத்துவோர் மற்றும் பயன்பாட்டு அடிப்படையில் இருக்கவேண்டும் என்ற உணர்வு, ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு உள்ள மொழிபெயர்ப்பை மேற்கொள்ளுவது இந்திய மொழிகளுக்கு இடையிலான இயந்திர மொழிபெயர்ப்பை மேற்கொள்வதைக் காட்டிலும் சிறந்தது என்ற தீர்மானத்திற்கு வழிவகுத்தது. மூன்றாவது ஒரு குறிப்பிட்ட பொருண்மைக்களத்தை அடிப்படையாகக் கொண்ட ஒழுங்குமுறைகள் (domain specitic systems) திறந்த ஒழுங்குமுறைகளைக் காட்டிலும் (open ended systems) உருவாக்குவதற்கு எளியது மற்றும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பு பயன்படுத்துவோர்க்கு சிக்கனமானதாகவும்

இருக்கிறது. இறுதியாக வேறுப்பட்ட மொழிகள் ஒற்றுமையுள்ள மொழிகளைக் காட்டிலும் ஆய்வு அடிப்படையில் கூடுதல் அறை கூவல்களை எதிர்கொள்ளும்.

சில முக்கியத்திட்டக் கருத்தல்கள் 90 விழுக்காடு வேலையை இயந்திரத்தாலும் வேலையை மனித முன் திருத்தத்தாலும் (post editing) மேற்கொள்ளும் படிக்கு மொழிப் பெயர்ப்பிற்கு ஒரு பயன்பாட்டு கருவியைத் தருவது என்ற நோக்கம் அடிப்படையில் அமைந்தது, இந்த அமைப்பொழுங்கு படிப்படியாகச் சிக்கலான சூழல்களைக் கையாளும்படிக்கு வளர இயலும், பொத்துமான உரை உருவாக்கும் பகுதிகளின் (text generator modules) இணைப்பால் ஆங்கிலத்திலிருந்து பெரும்பாலான இந்திய மொழிகளுக்கு மொழிப்பெயர்ப்பு ஒரு சீரான இயங்கு முறையைக் கொண்டிருந்தது. மேலும் அதன் பயன்பாட்டிற்கும் விரிவாகத்திற்கும் வசதி செய்ய மனிதனால் இயக்கப்பட்ட மனிதன் - இயந்திர இடைமுகம் கொண்டிருந்தது.

இது ஒரு குழும இந்திய மொழிகளுக்கு (target languages) பயன்படுமாறு போலி இலக்குமொழியை உருவாக்கும் (Pseudo target language) சூழல் வரையறையற்ற இலக்கணம் போன்ற அமைப்புடன் கூடிய அமைப்பொழுங்கை இலக்காகக் கொண்ட விதி அடிப்படையிலான ஒழுங்குமுறையாகும். பெருந்தரவு ஆய்வின் மூலம் கிடைக்கப்பெற்ற ஒரு குழும விதிகள் போலி மூலமொழிக்கு நகர்வு விதிகளைப் பயன்படுத்தும் சாத்தியமான உறுப்புகளைக் கண்டு கொள்வதற்குப் பயன்படுத்தப்படுகிறது. போலி இலக்குமொழியைப் பயன்படுத்தும் கருத்து, இடைமொழி அணுகுமுறையைப் பயன்படுத்துவதற்குச் சமமான நன்மையைப் பெற அமைப்புசார் ஒற்றுமையை முதன்மையாகப் பயன்படுத்துவதாகும். போலி இலக்குமொழி, இடைமொழி அணுகுமுறையில் பயன்படுத்தப்படும் இடைப்பட்ட மொழி அல்ல இங்கு பொருண்மை உருப்படுத்தத்தை உருவாக்க எந்த முயற்சியும் மேற்கொள்ளப்படவில்லை.

இந்திய மொழிகள் சுதந்திர சொல்குழும நிரலைக் கொண்டு வினையில் முடியும் பண்புடையனவாகும். அமைப்பு ஒற்றுமையின் வழி இந்திய மொழிகளை அவற்றின் மூலம் அடிப்படையில் நான்கு பெரிய குழுமங்களாக வகைப்படுத்தலாம்:

1. இந்திய – ஆரிய மொழிக்குடும்பம் (Indo-Aryan family) (இந்தி, வங்காளம் அஸ்ஸாமிஸ், பஞ்சாபி, மராத்தி, ஒரியா குஜாத்தி போன்றன)
2. திராவிட மொழிக் குடும்பம் (Dravidian Family) (தமிழ், தெலுங்கு, கன்னடம், மலையாளம் போன்ற)
3. ஆஸ்ட்ரோ – ஆசிய மொழிக் குடும்பம் (Astro-Asian family),
4. திபெத்திய பர்மிய மொழிக் குடும்பம் (Tibato-Burman Family).

ஒவ்வொரு குழுமங்களுக்குள்ளும் மொழிகள் உயர்ந்த அளவு அமைப்புப் பொருத்தத்தைக் காட்டுக்கின்றன. ஆங்கில பாரதி திட்டம் இந்த ஒற்றுமையைக் கூடுதல் அளவு இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கப்பயன்படுத்துகிறது.

இலக்குமொழியில் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கு வேண்டி பல பொருண்மை அடையாளங்கள் பயன்படுத்தப்படுகின்றன. தீர்க்கப்படாத பொருண்மை மயக்கங்களின் மாற்றுப் பொருள்கள் போலி இலக்குமொழியில் தக்க வைக்கப்படுகிறது. ஒவ்வொரு இலக்குமொழிக்கும் உரை உருவாக்கப் பகுதி போலி இலக்குமொழியை இலக்குமொழிக்கு மாற்றுகிறது. ஓவ்வொரு இலக்குமொழிக்கும் திருத்துவான் (corrector) பயன்படுத்தப்படுகிறது. இறுதியாக மனித இயக்கத்தால் தூண்டப்பட்டபின் திருத்தும் தொகுதி (Post editing package) இறுதி திருத்தங்களைச் செய்வத்றகுப் பயன்படுத்தப்படுகிறது. பின்திருத்தி இலக்குமொழியை மட்டும் அறிந்தால் போதுமானது.

ஆங்கில பாரதி திட்டம் சின்ஹா (Sinha, 1993) என்பவரால் 1991-ல் உருவாக்கப்பட்டது. இத்திட்டத்தில் ஆங்கிலத்திலிருந்து இந்தி மற்றும் தெலுங்கு மொழிகளுக்குச் செயல்படக்கூடிய மூலமுன்மாதிரி மொழிபெயர்ப்பு (Functional Prototype) ஒழுங்குமுறை உருவாக்கப்பட்டது. பின்வரும் திட்டவரைபடம் இந்த மொழி பெயர்ப்பு ஒழுங்குமுறையை விளக்கும். பின்வருவன ஆங்கில பாரதி மொழிபெயர்ப்புத் திட்டத்தின் முக்கியக்கூறுகளாகும்.

1. விதி அடிப்படை

விதி அடிப்படை (Rule base) ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு வாக்கியங்களின் அமைப்புகளைப் பொருத்துவதற்கான விதிகளைக் கொண்டிருக்கின்றது. இந்த ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு அமைப்பு ஒழுங்கு மாற்றங்களின் தரவு மையம் மொழிபெயர்களுக்கு அமைப்பு ஒழுங்கு மாற்றங்கள் தரவு மையம் மொழிபெயர்க்கவேண்டிய வாக்கியத்திற்குகாகக் கிளை அமைப்பைப் பெறும் வேலையைப் புறக்கணித்து புறக்கிளைகளிலிருந்து புறக்கிளை மாற்றத்தை செய்யும் வேலைக்குப் பணிக்கப்படுகிறது. வாக்கியங்கள் பொருண்மையைக் கண்டு பிடிக்க புற அமைப்பொழுங்கைப் பயன்படுத்தும் கருத்து, மொழியியலில் மிகப் பழமையானதாகும். இந்த அணுகுமுறை எளிதானது என்றாலும் இத்திட்டம் ஒரு மொழியின் புற அமைப்பு ஒழுங்களின் தனிப்பட்ட தன்மைகளைக் கண்டுபிடிக்கச் செயலூக்கமுடையதாகும். ஆங்கில பாரதியில் பயன்படுத்தப்படும் அணுகுமுறை சாம்ஸகியின் தொடரமைப்பு இலக்கணத்திலிருந்தும் சொல்சார்- செயல்பாடு சார் இலக்கணத்தின் (Lexical functional Grammer) அமைப்பிலிருந்தும் (C-structure) அதிகமாக எடுத்தாண்டுள்ளது. விதி அடிப்படை (rule base) என்று கூறப்படும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கான அமைப்பு மாற்ற விதிகளின் தரவு மூலம் ஆங்கில பாரதி ஒழுங்குமுறையில் மையப் பகுதியாக அமைகிறது. இது

ஆங்கிலத்திலிருந்து மொழிபெயர்கும் போது தொடரியலில் ஏற்படும் முக்கிய மாற்றங்களைக் கவனித்துக்கொள்கிறது. முன்னர் கூறியப்படி இந்திய மொழிகளுக்குப் பொது விதி அடிப்படையை உருவாக்கி ஆங்கில பாரதி ஆங்கிலத்திலிருந்து மொழிபெயர்க்கும் போது சிறந்த நன்மையை வெளிப்படுத்துகிறது.

2. அர்த்த மயக்க நீக்கி

அர்த்த மயக்கநீக்கி (Sense disambiguator) மூல மொழியின் ஒவ்வொரு சொல்லுக்கும் சரியான அர்த்ததை தேர்ந்தெடுக்கப் பொறுப்புள்ளதாகும். இங்கு அர்த்த மயக்கங்க நிக்கம் மூல மொழியின் உரையில் மட்டுமே நிகழ்த்தப் பெறுகின்றது. ஆங்கில பாரதியில் பயன்படுத்தப்படும் இவ்வணுகுமுறையை விதியால் விதி-பொருகோள் (Rule-by-rule Semantic Interpretation) எனக் கூறலாம். ஒரு தொடரியல் விதி பயன்படுத்தப்படும் ஒவ்வொரு நேரத்திலும் பொருள்கோளி (Semantic Intepreter) பயன்படுத்தப்படுகின்றது.

3. இலக்கு உரை உருவாக்கி

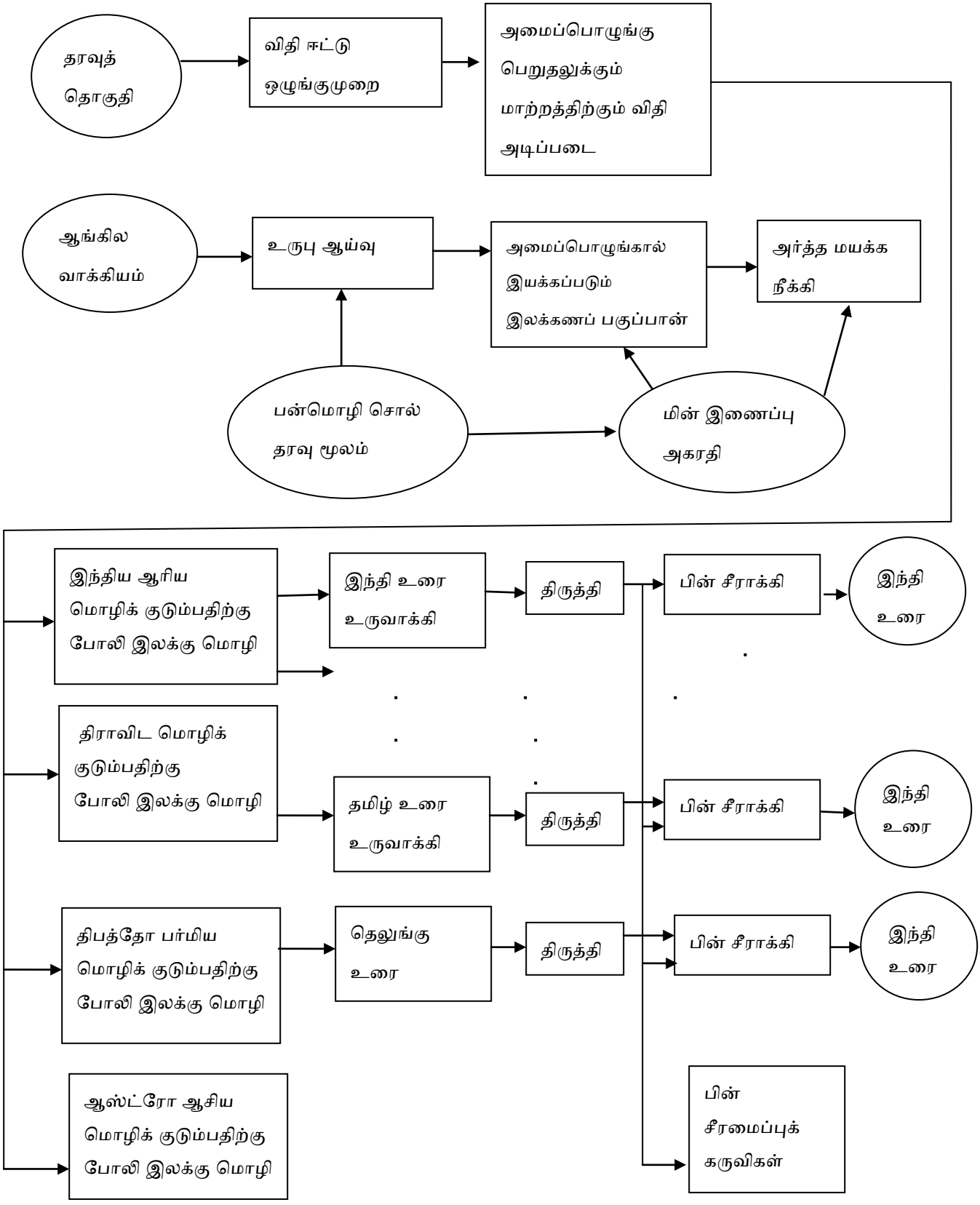
இலக்கு உரை உருவாக்கி (Target text generator) மொழிபெயர்ப்பு ஒழுங்குமுறையின் இறுதியாக அமைகின்றது. இதன் செயல்பாடு பொறுத்தமுறும் மூல மொழிகளிலிருந்து மொழிபெயர்க்கப்பட்ட வெளியீட்டை உருவாக்குவது ஆகும். இது ஆங்கில பாரதியின் முந்தைய நிலைகளின் உருவாக்கப்பட்ட இடைப்பட்ட வடிவத்தை உள்ளீடாக ஏற்றுக் கொள்கின்றது. இவ்வேலையை இயற்கை மொழி உருவாக்கம் (Natural Language Generation) என்று அழைக்கப்படும் வேலையிலிருந்து வேறுபட்டவையாகும்; இவ்வர்த்ததில் பிந்தையது எனக் கூறவேண்டும் என்பதுடன் திட்ட நிலை (Strategic level) எவ்வாறு கூற வேண்டும் என்பதையும் நடவடிக்கை நிலை (Tactical Level) தீர்மானிக்க வேண்டும். ஒரு விதி அடிப்படை மற்றும் அர்த்த மயக்க நீக்கி இவற்றைப் பயன்படுத்தும் வேறுபட்ட உரை உருவாக்கிகளைக் கொண்டு பல இலக்கு மொழிகளுக்கும் பொது இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை அமைப்பு கிடைக்கப்பெறுகின்றது.

4. பன்மொழி அகராதி

பன்மொழி அகராதி (Multilingual Dictionary) தொடரியல் வகைப்பாடுகள், சாத்தியமான அர்த்தங்கள், அர்த்தங்களின் மயக்கத்தை நீக்கும் முக்கிய கூறுகள் (Keys) இவற்றை உள்ளடக்கிய ஆங்கிலச் சொற்களுக்கும் இணையாக மூல மொழிச் சொற்களுக்குப் பல விளக்கங்களைக் கொண்டிருக்கும்.

5. விதி அடிப்படை ஈட்டி

விதி அடிப்படை ஈட்டி (rule base acquirer) மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு விதி அடிப்படையை உருவாக்குகிறது.



10.7.1.4.அனுபாரதி

1995-இல் சின்ஹா அனுபாரதி வழிமுறையை உருவாக்கினார். இது எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு உபாயத்தைப் பின்பற்றியது. இந்த வழிமுறை இந்தியிலிருந்து ஆங்கிலத்திற்கு மொழிபெயர்க்கப் பயன்படுத்தப்பட்டது. அனுபாரதியின் அணுகுமுறை ஒற்றுமையுள்ள மொழிகளுக்கிடையில், எடுத்துக்காட்டாக இந்திய மொழிகளுக்கு இடையில் இயந்திர மொழிபெயர்ப்பு செய்யத் திறமையாகச் செயல்பட்டது. அம்மாதிரியான நேர்வுகளில் சொல்வரிசை ஒன்றாக இருக்கும் மற்றும் பொருத்தங்களை நிறுவதற்கு சுட்டிக்காட்டிகள் தேவையில்லை.

ஆங்கிலபாரதி, அனுபாரதி ஆகிய இரண்டு இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளும் அதன் தொடக்க கருத்துருவாக்கத்திலிருந்து கணிசமான மாற்றங்கள் அடைந்தன. 2004-இல் மொழிபெயப்பு ஒழுங்குமுறையின் முன்னேற்றத்தின் இரண்டாவது கட்டம் வெளியிடப்பட்டது. இது முந்தைய கட்டமைப்பின் குறைகளை நிவர்த்தி செய்தது. இவை ஆங்கிலபாரதி II என்றும் அனுபாரதி II என்றும் பெயரிடப்பட்டன. இவ்விரு இயந்திரபெயர்ப்பு ஒழுங்குமுறைகளும் வேறுபட்ட முன்னுதாரணங்களின் கலப்பாக்கத்தின் மாறுபடும் அளபுகளில் கலப்பாக்கம் செய்யப்பட்டுள்ளன.

10.7.1.5.ஆங்கிலபாரதி II (2004)

ஆங்கிலபாரதி-II (Sinha, 2004) முந்தைய மொழிபெயர்ப்பு ஒழுங்குமுறையின் கட்டமைப்பின் பல குறைபாடுகளை நீக்கியுள்ளது. இது கலப்பிற்காகப் பொது எடுத்துக்காட்டு அடிப்படையையும் (Generalized Eample-base (GEB)) மூல எடுத்துக்காட்டு அடிப்படையையும் (Raw Example-base (REB)) பயன்படுத்துகின்றது. உருவாக்கும் கட்டத்தில் விதி-அடிப்படையில் மாற்றுவது சிக்கலானது என்றும் ஊகிக்க இயலாத விளைவுகளைத் தரும் என்றும் தெரியவந்தபோது எடுத்துக்காட்டு அடிப்படை மேம்படுத்தப்பட்டு ஊடாட்டமாக முன்னேற்றமடைந்தது. உண்மையான பயன்பாட்டின் போது, ஒழுங்குமுறை விதி அடிப்படையை தருவிப்பதற்குமுன் முதலில் பொது எடுத்துக்காட்டு அடிப்படைக்கும் மூல எடுத்துக்காட்டு அடிப்படைக்கும் பொருத்தத்தை முயற்கிக்கும். ஆங்கிலபாரதி- II-இல் தானியக்க முன் திருத்ததிற்கும் (pre-editiong) பொழிப்புரைக்கும் (paraphrasing) பொதுமைப்படுத்தப்பட்ட மற்றும் கட்டுப்பட்ட பல்சொல் வெளிபடுத்தங்களுக்கும் (generalized and conditional expressions), பெயரிடப்பட்ட பொருள்களை (named entities) அறிந்துகொள்வதற்கும், பொருள்புலத் தனிப்பயனாக்க கருவிகளுக்கும் (domain customization tools), உட்படுத்தப்பட்ட பிழையாய்வுத்

தொகுதிக்கும் தானியக்கப் பின் திருத்தத்திற்கான புள்ளியியல்சார் மொழி மாதிரிக்கும் (statistical language model) ஏற்பாடு செய்யப்பட்டிருந்தது.

அனுபாரதி உத்தி அனுபாரதி II-இல் இந்தியை மூலமொழியாகக்கொண்டு பிறமொழிகளுக்கு மொழிபெயப்பதற்கு ஏதுவாக எடுத்துக்காட்டு அடிப்படையின் பொதுமையாக்கம் மூலமொழியைச் சாதிருக்கும் படி பொதுமைப்படுத்தப்பட்டுள்ளது. அனுபாரதி-II-இன் கட்டமைப்பின் மையம் ஒரு பொதுமையாக்கப்பட்ட படிநிலைசார் எடுத்துக்காட்டு அடிப்படையாகும். இந்தியிலிருந்து பிற இந்திய மொழிகளுக்கு எடுத்துக்காட்டு அடிப்படையின் உருவாக்கம் ஒற்றுமையிலாத மொழிகளுடன் ஒப்பிடும் போது மிக எளிமையானதாகும். ஒப்பீட்டளவில் இந்தி பிற எல்லா இந்திய மொழிகளைப் போன்று சுதந்திரமான சொல் வரிசையுள்ள மொழியாகும். உள்ளீடு செய்யப்படும் இந்தி வாக்கியம் சொல்வரிசை வேறுபாடுகளைக் கையளுவதற்காக ஆழமில்லா இலக்கணப் பகுப்பாய்வால் தரப்படுத்தப்பட்ட வடிவாக மாற்றப்படுகின்றது. தரப்படுத்தப்பட்ட இந்தி வாக்கியம் மேல்நிலை தரப்படுத்தப்பட்ட எடுத்துக்காட்டு அடிப்படையுடன் பொருத்தப்படும். பொருத்தம் கண்டுபிடிக்கபடாவிட்டால் ஒரு ஆழமில்லா தொடர்பகுப்பான் (shallow chunker) உள்ளீட்டு வாக்கியத்தை அலகுகளாகத் துண்டுபடுத்தப் பயன்படுத்தப்படும்; இத்துண்டுகள் படிநிலைசார் எடுத்துக்காட்டு-அடிப்படையுடன் பொருத்தப்படும். மொழிபெயர்ப்பு செய்யப்பட்ட தொடர்கள் (chunks) வாக்கிய நிலை எடுத்துக்காட்டு அடிப்படையுடன் பொருத்தத்தால் நிலைபடுத்தப்படும். பிழைஆய்வு தொகுதியும் புள்ளியியல் மொழி மாதிரியும் ஆங்கிலபாரதி II போன்று உட்படுத்தப்பட்டுள்ளன. மனிதப் பின்திருத்தம் முக்கியமாக இந்தியில் இருக்கிற அல்லது மதிப்பிட சிக்கலான அடைகொளி அடைகளைப் புகுத்த வேண்டி செய்யப்படுகின்றது.

10.7.1.6.மந்ரா இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை (MANTRA MT System)

மந்த்ரா (MAchiNe assisted TRAnslation tool (MANTRA)) ஆங்கிலப் பனுவலை இந்தியில் மொழிபெயர்க்கும். இது தனிப்பட்ட நிர்வாகம். அரசிதழ் அறிவிப்புகள், அலுவலகக் கட்டளைகள், அலுவலக அறிக்கைகள் மற்றும் சுற்றுக்கள் போன்ற குறிப்பிட்ட பொருண்மைகளங்களில் மொழிபெயர்ப்பு செய்யும். ஆங்கில மற்றும் இந்தி இலக்கணத்தை உருப்படுத்தம் செய்ய மந்ரா சொல்லாக்கப்பட்ட கிளையமைப்பு இணையும் இலக்கணத்தைப் (Lexicalized Tree Ajoining Grammar (LTAG)) பயன்பத்துகின்றது. இது கிளையமைப்பு இணையும் இலக்கணத்தை (Tree Ajoining Grammar (TAG)) பகுப்பாய்வு செய்யவும் ஆக்கம் செய்யவும் பயன்படுத்தும். மந்த்ரா

ஒழுங்குமுறைக்கு பனுவல் ஆவணங்கள், பேச்சு அறிதல் நிரல் அல்லது ஒளிஎழுத்துணரி தொகுப்புக்கள் வழியாக உள்ளீடு செய்யலாம்.

தொடக்கத்தில் மந்த்ரா ஒழுங்குமுறை நியமனக்கடிதம், அறிவிப்பு, மைய அரசால் வெளியிடப்படும் சுற்றறிக்கை போன்ற நிர்வாக ஆவணங்களின் மொழிபெயர்ப்புக்காகத் தொடங்கப்பட்டது. பின்னர் இவ்வொழுங்குமுறை நிதி, விவசாயம், உடல்நலப்பராமரிப்பு, தகவல் தொழில்நுட்பம், கல்வி ஆகியப் பொருண்மைக்களத்திற்கும் அரசாங்கப் பொருண்மைக்களத்தில் பொதுவான நோக்கச் செயல்பாட்டிற்கும் நீட்சிசெய்யப்பட்டது. மந்த்ரா தொழில்நுட்பதை அடிப்படையாகக் கொண்ட மந்த்ரா-ராஷ்ட்ரபாஷா என்ற திட்டம் சி-டாக்கால் உருவக்கப்பட்டது. (http://www.cdac.in/html/aai/mantra.asp).

பணித்துறை பொருண்மைக்களத்தில் வாக்கியக் கட்டுமானங்களை ஏற்று பகுத்தாய்வுசெய்து உருவாக்கம் செய்வதற்கு வேண்டி இலக்கணம் சிறப்பாக வடிவமைக்கப்பட்டது. இதுபோன்று அதன் பொருண்மைக் களங்களில் பயன்படுத்தப்படும் ஆங்கிலச் சொற்களின் பொருள்களைக் கையாளுவதற்காக அகராதி பொருத்தமாகக் கட்டுப்படுத்தப்பட்டுள்ளது.

மந்த்ரா-ராஷ்ட்ரபாஷா என்ற ஒழுங்குமுறை இந்தியப் பாராளுமன்றத்தின் மேல்சபையில் ராஜ்யசபா செயலத்திற்காக உருவாக்கப்பட்டது. இது பாராளுமன்ற நடவைக்கைகளை மொழிபெயர்கின்றது. தற்போது ஆங்கிலம்-பெங்காளி, ஆங்கிலம்-தெலுங்கு, ஆங்கிலம்-குஜராத்தி, இந்தி-ஆங்கிலம், இந்தி-பெங்காளி, இந்தி-மராத்தி ஆகிய மொழி இணைகளுக்குத் தொடங்கப்பட்டுள்ளது.

10.7.1.7.அனுவாதக் மொழிபெயர்ப்பு ஒழுங்குமுறை

சூப்பர் இன்ஃபோசாஃப்ட் ப்ரைவேட் லிமிடெட்(Super Infosoft Private Limited), தில்லி அனுவாதக் 5.0 என்ற ஆங்கிலம்-இந்தி இயந்திர மொழிபெயர்புக் கருவியை உருவாக்கியது. இது முன் திருத்தத்திற்கு ஆதரவு தரும். இது அதிகாரம், முறைசார்ந்தவை, விவசாயம், மொழியியல், தொழிநுட்பம், நிருவாகம் போன்ற சிறப்பான பொருண்மைகளங்களுக்கு உட்படுத்தப்பட்ட அகராதியைக் கொண்டுள்ளது. ஆங்கிலச்சொல்லுக்கு இந்திப் பொருள் இல்லை என்றால் ஒலிபெயர்ப்பு வசதி தரப்பட்டுள்ளது. மென்பொருள் எந்த விண்டோ குடும்பத்தின் எந்த இயக்கமுறையிலும் (operating system) வேலைசெய்யும்.

10.7.1.8.உலகளாவிய வலைப்பின்னல் மொழி அடிபடையிலான ஆங்கில இந்தி மொழிபெயர்ப்பு ஒழுங்குமுறை (UNL-based English-Hindi MT system)

பாம்பேயிலுள்ள இந்திய தொழில்நுட்ப நிறுவனம் உலகளாவிய வலைப்பின்னல் மொழியைப் பயன்படுத்தி (Universal Networking Language (UNL) ஆங்கிலம்-இந்தி மொழிபெயர்ப்புக்காக ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்கியது. யுஎன்எல் ஐக்கிய நாடுகளின் பல்கலைக்கழகத்தின் உலகத்திட்டம் ஆகும். இதன் நோக்கம் எல்லா முக்கிய மனித மொழிகளுக்கும் ஒரு இடைமொழியை உருவாக்குவதாகும். பாம்பேயிலுள்ள இந்திய தொழில்நுட்ப நிறுவனம் யுஎன்எல்-இன் இந்திய உறுப்பினராகும். (இந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை பின்வரும் வலைதளத்தில் பயன்படுத்தலாம்: http://www.cfilt.iitb.ac.in/machine-translation/eng-hindi-mt. மேலும் இரண்டு செய்துகாட்டும் ஒழுக்குமுறைகளையும் இந்த வலைதளத்திலிருந்து பயன்படுத்தலாம்: ஒன்று இந்தி-யுஎன்எல் மற்றொன்று யுஎன்எல்-இந்தி மாற்றம். யுஎன்எல் அடிப்படையிலான ஆங்கிலம்-மராத்தி, ஆங்கிலம்-பெங்காளி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் முயற்சிக்கப்பட்டு வருகின்றன.

10.7.1.9.மாத்ரா இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை

தேசிய மென்பொருள் தொழிநுட்ப மையத்தின் (National Centre for Software Technology (NCST), மும்பையிலுள்ள (தற்போது சிடாக், மும்பை) அறிவு அடிப்படை கணினி ஒழுங்குமுறைகளின் (Knowledge Based Computer Systems) பிரிவின் இயற்கைமொழி ஆய்வுக் குழு மாத்ராவை உருவாக்கியது. இது ஆங்கிலத்திலிருந்து இந்திக்கு மனித உதவியுடனான மற்றல்-அடிப்படை மொழிபெயர்ப்பு ஒழுங்குமுறையாகும். இது டிடியல்-ஆல் ஆதரிக்கப்பட்ட ஆய்வாகும். மாத்ரா அகராதி மற்றும் அணுகுமுறை பொது நோக்கத்திற்கானதாகும். இந்த ஒழுங்குமுறை செய்தி, ஆண்டறிக்கைகள், தொழில்நுட்ப தொடர்கள் ஆகிய பொருண்மைகளங்களை மொழிபெயர்க்கப் பயன்படுத்தப்பட்டது.

ஆங்கிலச் செய்திக் கதைகளை இந்தியில் மொழிபெயர்ப்பதற்காக வாக்யா (Vaakya) என்ற கலப்பு அணுகுமுறை ஒழுங்குமுறை என்சிஎஸ்டி, பாம்பே-இல் உருவாக்கப்பட்டது. இந்த ஒழுங்குமுறை ஒரு தனி வினை வாக்கியத்தைக் கையாள இயலும். மூலமுன்மாதிரி வாக்கியா (Protype Vaakya) ஒழுங்குமுறை பின்னர் ஆங்கிலச் செய்திக் கதைகளை இந்திக்கு மொழிபெயர்க்கும் சேவையைக் கொடுக்கும்படி மேம்படுத்தப்பட்டு மாற்றியமைக்கப்பட்டுள்ளது.

10.7.1.10.ஆங்கில-கன்னடா இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை

ஹைதராபாத் பல்கலைகழகத்தின் கணினி மற்றும் தகவல் அறிவியல் துறை (Computer and Information Sciences) உலகலளாவிய எச்சத்தொடர் அமைப்பு இலக்கணம் (Universal

Clause Structure Grammar (UCSG)) வடிவவாதத்தைப் பயன்படுத்தி ஆங்கிலம்-கனடா இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்கியது. இது மாற்றல் அடிப்படையில் அமைந்த அணுகுமுறையாகும். இது அரசு சுற்றுகள் பொருண்மைக்களத்தை மொழிபெயர்க்க பயன்படுத்தப்படும் இந்த ஒழுங்குமுறை கர்நாடக அரசின் நிதியுதவியால் மேற்கொள்ளப்பட்டது. (http://www.languagetechnolies.ac.in/lerc/mat/mat.htm). இது ஆங்கிலம்-தெலுங்கு மொழிபெயர்ப்பு ஒழுங்குமுறையாயகவும் நீட்சி செய்யப்பட்டுள்ளது.

10.7.1.11. இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் ஒப்பீடு

சில இந்திய இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் ஒருவகையான ஒப்பீட்டைப் பின்வரும் அட்டவணை வெளிப்படுத்தும்.

இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள்	அணுகுமுறை	மொழி இணைகள்	பண்புக்கூறுகள்
அனுசாரக	நேரடி	பெங்காளி, கன்னடா, மராத்தி, பஞ்சாபி, தெலுங்கு இவற்றிலிருந்து இந்தி	பாணினிய இலக்கணதைப் பயன்படுதுகின்றது; மூலமொழிக்கும் இலக்குமொழிக்கும் இடையில் இடஞ்சார் சொற்களைப் பொருந்துகின்றது
இந்தியிலிருந்து பஞ்சாபி இயந்திரமொழி பெயர்ப்பு ஒழுங்குமுறை	நேரடி	இந்தியிலிருந்து பஞ்சாபி	சொல்லிலிருந்து சொல் நேரடி மொழிபெயர்ப்பு; உருபனியல் பகுப்பாய்வு, சொற்பொருள் மயக்கநீக்கம்,
மந்த்ரா	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து இந்தி, குஜராத்தி, தெலுகு; இந்தியிலிருந்து ஆங்கிலம், பெங்காளி, மராத்தி	கிளையமைப்பு இணைப்பு இலக்கண வடிவமைப்பு

மாத்ரா	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து இந்தி	மனித உதவியுடன் கூடிய மொழிபெயர்ப்பு; விதி அடிப்படையையும் பட்டறிவையும் பயன்பத்துகின்றது
சக்தி	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து இந்தி, மராத்தி, தெலுகு	விதி அடிப்படையையும் புள்ளியியல் அணுகுமுறையையும் இணைத்துச் செயல்படுகின்றது
அனுபாத்	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து பெங்காளி	கலப்பு ஒழுங்குமுறை; சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கு என்-கிராம் அணுகுமுறையைப் பயன்படுத்துகின்றது; வாக்கிய நிலையில் வேலை செய்கின்றது
ஆங்கிலம்-கன்னடா இயந்திரமொழி பெயர்ப்பு ஒழுங்குமுறை	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து கன்னடா	உலகளாவிய எச்சத்தொடர் அமைப்பு இலக்கண வடிவமைப்பை பயன்படுத்துகின்றது.
சம்பர்க்	மாற்றல் அடிப்படை	பஞ்சாபி-இந்தி, தெலுங்கு-தமிழ், இந்தி-உருது, இந்தி-தெலுங்கு	மொழிப் பகுப்பாய்விற்கு கணினிசார் பாணினி இலக்கண அணுகுமுறையைப் பயன்படுத்துகின்றது; இதை இயந்திரம் கற்றலுடன் இணக்கின்றது
ஆங்கிலபாரதி	இடைமொழி சார்	ஆங்கிலம்-இந்தி	அடிக்கடி நேரிடும் பெயர்த்தொடர்களுக்கும் வினைத்தொடர்களுக்கும் மொழிபெயர்ப்பைப் பெறுவதற்காக விதி அடிப்படை, எடுத்துக்காட்டு

			அடிப்படை, புள்ளியியல் இவற்றைப் பயன்பத்துகிறது
உலகளாவிய இயற்கை மொழி ஆங்கில இந்தி மொழிபெயர்ப்பு ஒழுங்குமுறை	இடைமொழி சார்	ஆங்கிலம்-இந்தி	உலகளாவிய இயற்கை மொழியை இடைமொழியாகப் பயன்படுத்துகின்றது
ஆங்கிலம் இந்தி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	புள்ளியியல் சார் இயந்திர மொழிபெயர்ப்பு	ஆங்கிலம்-இந்தி	விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பையும் தொடர் அடிப்படையிலான புள்ளியியல் இயந்திர மொழிபெயர்ப்பையும் இணைக்கின்றதும்
ஆங்கிலம் மலையாளம் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	புள்ளியியல் சார் இயந்திர மொழிபெயர்ப்பு	ஆங்கிலம்-மலையாளம்	ஒருமொழிய மலையளத் தரவுத்தொகுதியையும் புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு இருமொழிய ஆங்கிலம்/மலையாளம் தரவுத்தொகுதியையும் பயன்படுத்துகின்றது.
வாசானுபாத	எடுத்துக்காட்டுசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	இருமொழிய பெங்காளி அசாமீஸ்	முன் செயற்பாங்கும் பின்செயற்பாங்கு செயல்பாடு, நீண்ட வாக்கியங்கள் நிறுத்தற்குறிகள் வரும் இடத்தில் துண்டுபடுத்தப்படும்; பொருந்தாத விளைவுகளுக்குப் பின்தடம்பற்றும்
அனுபாரதி	எடுத்துக்காட்டுசார் இயந்திர மொழிபெயர்	ஆங்கிலம்-இந்தி	கலப்பு எடுத்துக்காட்டு அடிப்படையிலான ஒழுங்குமுறையான இது அமைப்பொழுங்கு

	ப்பு ஒழுங்குமுறை		அடிப்படையையும் எடுத்துக்காட்டு அடிப்படை அணுகுமுறையையும் தன்னுடன் இணைத்துக்கொள்ளும்.
சிவா	எடுத்துக்காட்டுசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	ஆங்கிலம்-இந்தி	மொழியியல் தகவல்களை ஊகிப்பதற்காக மொழியியல் விதிகளையும் புள்ளியியல் அணுகுமுறையும் பயன்படுத்தும்.

10.7.2.தமிழ் சார் இயந்திர மொழிப்பெயர்ப்புகள்

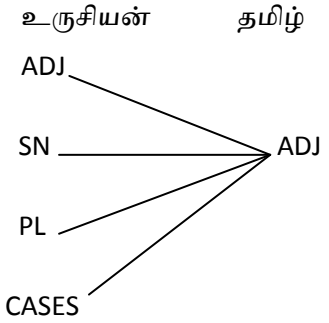
தமிழ்சார் இயந்திர மொழிபெயர்ப்புகள் பல மேற்கொள்ளப்பட்டு தமிழ் நாட்டில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சிக்கு வித்திட்டன.

10.7.2.1 தமிழ் உருஷ்யன் மொழிப் பெயர்ப்புத்திட்டம்

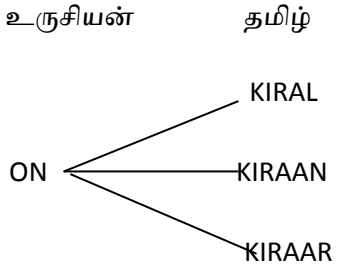
1985-ஆம் தமிழ் பல்கலைக்கழகத்தில் (Tumtst Tamil University Machine Inaslation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது (Chellammuthu et al 1984, Chellemuthu 2002). இதன் மூலம் ருஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழிப்பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக்கழகம் கண்ணி அறிவியல்துறை, மொழியியல் துறை மற்றும் மொழிப் பொயர்ப்புத் துறை ஆகியவைகளின் கூட்டு முயற்சியில் உருவானதாகும்.

உருசியன் தமிழ் மொழிபெயர்ப்பிற்கான படிநிலைகளின் விளக்கம் கீழே தரப்பட்டுள்ளன. மூலமொழி உருசியனை L1 என்றும் தமிழ் மொழியை L2 என்றும் குறிப்பிடுகின்றன. முதற்கட்டமாக மூலமொழியில் உள்ள ஒரு செய்தி ஒலிபெயர்ப்பு செய்யப்பட்டு கணிப்பொறிக்கு உள்ளீடு செய்கின்றன. இரண்டாம் கட்ட நிலையில் மூல மொழியில் உள்ள செய்தி பகுப்பாய்வு செய்யப்படுகிறது. இந்நிலையில் சொற்களின் இலக்கண கூறுகள் கண்டறியப்படுகின்றன. இதற்கு ருஷ்யன் - தமிழ் அகராதிகள் உதவுகிறது. இப்பகுப்பாய்வு மூலம் ருஷ்ய மொழியில் உள்ள செய்திகள் NP, VP, NN, ADF என பாகுபடுத்தப்படுகின்றன. மூன்றாம் கட்ட நிலையில் பகுப்பாய்வு செய்யப்பட்டுள்ள மூலமொழியின் வார்த்தையில் உள்ள சொற்களின் படிவம், வாக்கிய அமைப்பு போன்றவைகளுக்கு இலக்கு மொழியில் நிகராக உள்ள சொற்களின் வடிவம், வாக்கிய அமைப்பு ஆகியன மாற்றம் (Transfer) என்ற அமைப்பின் மூலம் கண்டறியப்பட்டு மொழி மாற்றம் செய்யப்படுகின்றன. இந்நிலையில் இவ்விரு மொழிகளுகட்கும் இடையே உள்ள

முரண்பாட்டு அறிவு (Contrast Knoweldge) பயன்படுத்தப்படுகின்றன. கருத்தாடலைப் பொருத்தவரை இவ்விரு மொழிகளுக்கு இடையே ஒன்றுக்கு பல (One-to-Many) உறவு உள்ளது. பெயரடைச்சொல்லை பொறுத்தவரையில் இவ்விரு மொழிகளுக்கு இடையே (many-to-One) உறவு உள்ளது. இவ்வுறவினை கீழ்கண்ட படம் விளக்குகிறது.



மேலும் இவ்விரு மொழிகளுக்கு இடையே உள்ள வேற்றுமை உருபுகள் ஒன்றுகுப் பல (one-to-many) என்ற உறவைக்காட்டுகின்றன.



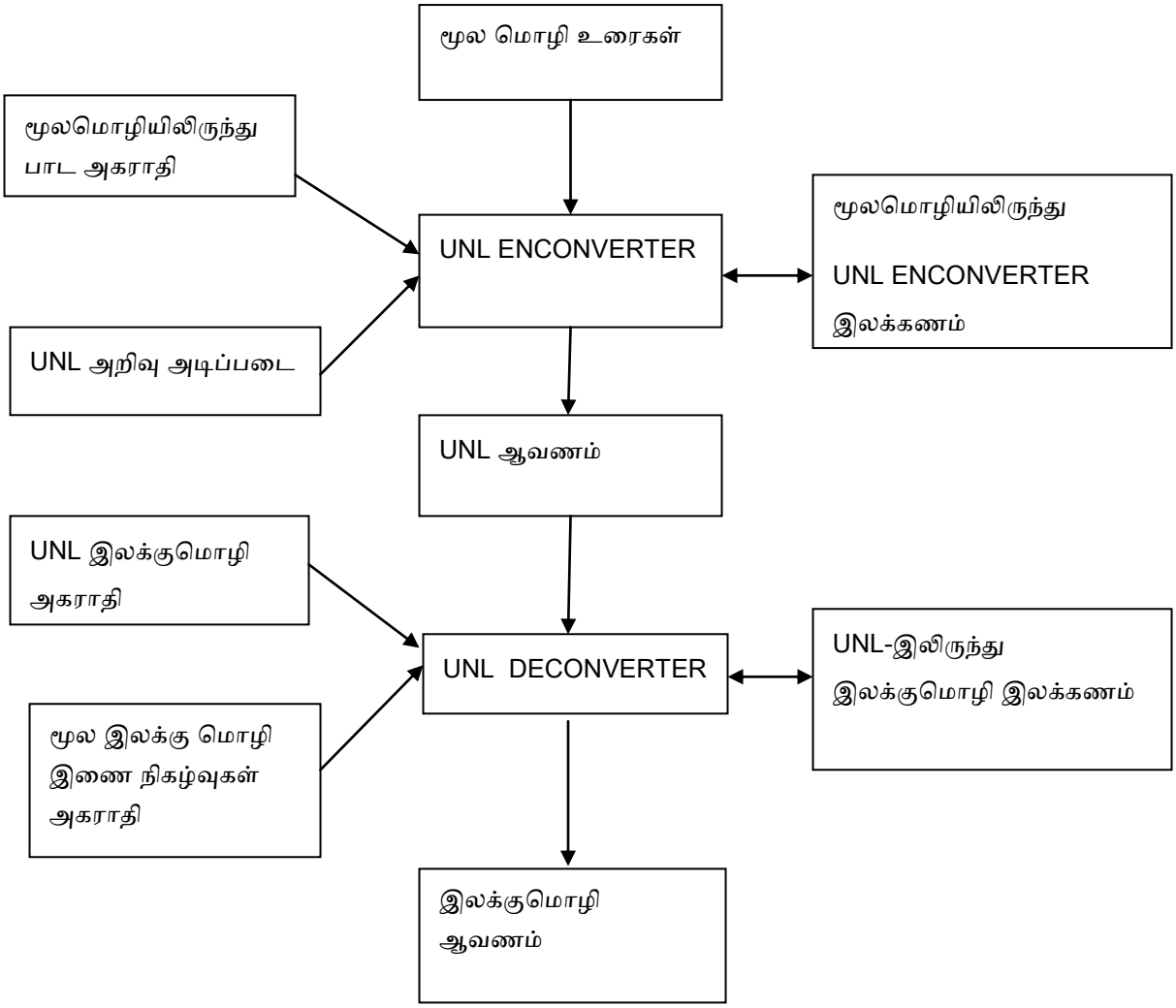
இதுபோன்ற இவ்விரு மொழிகளுக்கு இடையே உள்ள வேறுபாடுகள் விதிமுறைகளாகவும் வழிமுறைகளாகவும் மாற்றம் ((Transfer)) என்ற அமைப்பின் கீழ் உள்ளன. பகுப்பாய்வு செய்யப்பட்ட மூலமொழி செய்திகள் மாற்ற (Transfer) அமைப்பில் உள்ளீடாகச் செல்கிறது. இந்நிலையில் மூலமொழி அமைப்பிற்கு நிகரான இலக்கு மொழி அமைப்பிற்குச் செய்தி மொழிமாற்றம் செய்யப்படுகிறது. நான்காவது கட்ட நிலையில் மொழி மாற்றம் செய்யப்பட்டுள்ள இலக்கு மொழி கட்டமைப்புகள் இலக்கு மொழியின் தன்மைக்கேற்ப அமைவதில்லை. இவ்வாறு கிடைக்கப்பெற்ற இடைநிலை மொழிமாற்ற அமைப்புகளை இலக்கு மொழியில் தன்மைக்கேற்ப உருவாக்கித் தருவது தான் உருவாக்கம் (Genaration) என்ற பகுதியின் பணியாகும். இந்நிலையில் மொழிமாற்றம் செய்யப்பட்ட செய்திகள் உள்ளன. இச்செய்தியை மூலமொழியின்

வரிவடிவத்திலே அல்லது ஒலிபெயாப்பு வடிவத்திலே தரும் பணியை வெளியீடு (Output) என்ற அமைப்பு தருகிறது. மொழிபெயர்ப்பு செய்யப்பட்ட செய்தியைக் காட்சித் திரையில் காணலாம். தேவைப்படின் அச்சுப்பொறி வழி அச்சிட்டுக்கொள்ளலாம். உருசியன்-தமிழ் மொழிபெயர்ப்பைப் பொருத்தவரையில் வெளியீட்டை தமிழ் எழுத்துக்களில் காண்பதற்கும், அச்சிட்டுக் கொள்வதற்கும் வசதிகள் உள்ளன. அதுபோல் ஒலிபெயர்ப்பு முறையிலும் விடையைப் பெறலாம்.

10.7.2..2.உலக வலைப்பின்னல் மொழி – தமிழுக்கான இடைமொழி இயந்திர மொழிப்பெயர்ப்பு

இந்த அணுகுமுறை சென்னையில் உள்ள அண்ணா பல்கலைக்கழகத்தில் கணிப்பொறி மொழியியல் துறையில் மேற்கொள்ளப்பட்டது. (Dhanabalan & Geetha, 2004). இந்த அணுகு முறையில் இடையீட்டு மொழியின் உருப்படுத்தம் பயன்படுத்தப்படுகிறது. மூலமொழி அகராதியையும் இலக்கணத் தகவல்களையும் பயன்படுத்தி இடைமொழி உருப்படுத்தம் மூலமொழியிலிருந்து இலக்கு மொழிக்கு மொழிப்பெயர்ப்பு செய்ய பயன்படுத்தப்படும் இடைமொழி அமைப்பு உலகளாவிய வலைப்பின்னல் மொழி (Universal Networking Language (UNL)) ஆகும். இம்மொழி இடைமொழி அமைப்பாகத் திட்டமிடப்பட்டு மூலமொழியிலிருந்து இலக்கு மொழிக்கும் இலக்கு மொழியிலிருந்து மூலமொழிக்கும் தானியங்கி இயந்திர மொழிபெயர்ப்புயைச் சாத்தியமாக்குகின்றது. UNL-ஐ பயன்படுத்திய மொழிபெயர்ப்பு ஒழுங்குமுறை மூலமொழியிலிருந்து UNL-க்கு மாற Encoverter-ஐயும் UNL-இலிருந்து இலக்கு மொழிக்கு மாற Deconverter பயன்படுத்துகின்றது.

இந்த அணுகுமுறையின் திட்டவடிவம் கீழே தரப்பட்டுள்ளது.



மேற்கண்டவாறு படத்தில் காட்டியவாறு UNL அமைப்பு வேலை செய்து UNL மொழியிலிருந்து தமிழுக்கு மொழிபெயர்ப்பு செய்யும்.

10.7.2.3 ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்கும் திட்டம்

ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிபெயர்க்க ஒரு இயந்திர மொழி பெயர்ப்புக் கருவி உருவாக்க, காமாட்சி மற்றும் ராஜேந்திரன் (Kamachi Rajendran, 2004) ஒரு திட்டத்தை உருவாக்கி செயல்படுத்த முயன்றனர். தமிழில் முதுகலை மொழியியல் பாடத்திட்டத்திற்கு வேண்டிய மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்க வேண்டிய கட்டாயம் ஏற்பட்டதால் இத்தகைய ஆய்வு மேற்கொள்ளப்பட்டது. இவ்வாய்வு மொழியியல் புத்தகங்களை ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்க்க உதவும் ஒரு கருவி தயாரிப்பதை நோக்கமாக கொண்டது. இக்கருவி பின்வரும் நான்கு பகுதிகளைக் கொண்டது.

1. தமிழ் மொழி அமைப்புப் பகுத்தாய்வு

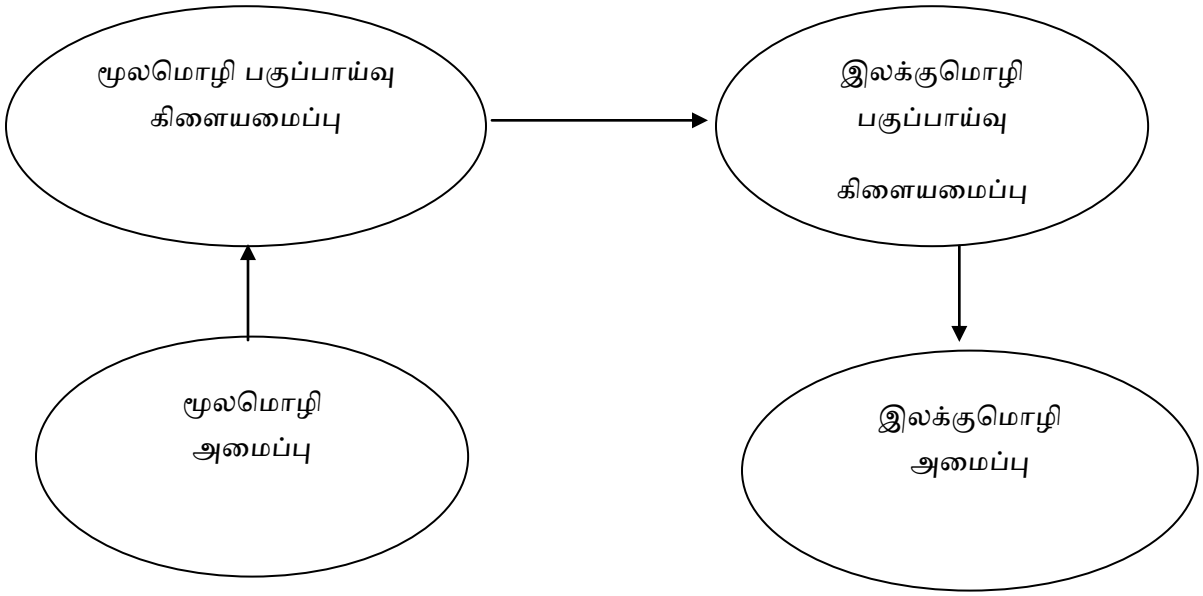
2. ஆங்கில மொழி அமைப்புப் பகுத்தாய்வு

3. ஆங்கில - தமிழ் மொழிபெயர்ப்பு அகராதி

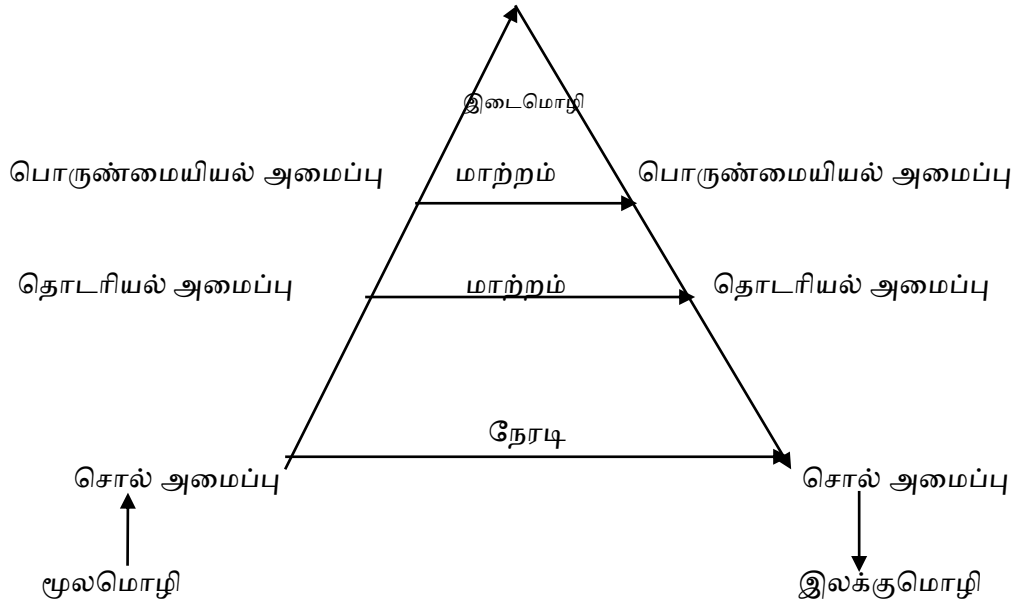
4. ஆங்கில – தமிழ்மொழி மாற்றமைப்பு விதிகள்

5. ஆங்கில மொழியமைப்பு உருவாக்கி

இந்த மொழிப்பெயர்ப்பு கருவியின் செயல்பாட்டை விளக்கும்



ஒரு ஒழுங்குமுறையை உருவாக்குவதில் ஏற்படும் சிக்கலின் நிலை மொழி பெயர்க்கப்படும் இரு மொழிகளின் ஒற்றுமை மற்றும் வேற்றுமைகளை பொறுத்து அமையும். இவ்வொற்றுமை வேற்றுமைகள் சொல்நிலையில் மொழி மாற்றம் செய்யப்பட வேண்டுமா? தொடர் நிலையில் மாற்றம் செய்யப்பட வேண்டுமா? பொருண்மை நிலையில் மாற்றம் செய்யப்பட வேண்டுமா என்பதைத் தீர்மானிக்கும் பின்வரும் படம் இச்சாத்தியங்களை வெளிப்படுத்தும்.



தமிழிலிருந்து பிற மொழிகளுக்கும் பிற மொழிகளிலிருந்து தமிழ்மொழிக்கும் இயந்திர உதவி மொழிபெயர்ப்பு ஒழுங்குமுறைகள் தமிழ்மொழியின் மீது பல நிலைகளிலான ஆய்வுகளை உள்ளடக்கும். ஒவ்வொரு நிலைக்கும் அதற்கான சிக்கல்கள் காணப்படும்.

10.7.2.4. ஏயுகேபிசி ஆய்வு நிறுவனத்தின் தமிழ்-இந்தி மொழிபெயர்ப்பு திட்டம்

ஐஐடி கான்பூரில் அனுசாரக இயந்திரமொழிபெயர்ப்புத்திட்டத்தின் ஒரு பகுதியாய் ஆயிரத்தித் தொள்ளாயிரத்தித் தொண்ணுறுகளில் இரசேந்திரனால் மேற்கொள்ளப்பட்ட இந்தி-தமிழ் மொழிபெயப்புத் திட்டம் இரண்டாயிரத்தின் தொடக்கத்தில் ஏயுகேபிசி ஆய்வு மையத்தால் ஏற்றெடுக்கப்பட்டு இராசேந்திரனின் மேற்பார்வையில் ஓரளவுக்கு இயந்திர மொழிபெயர்ப்பு வெளியீடு தரத்தக்கவிதத்தில் மேம்படுத்தப்பட்டது.

10.7.2.5. இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்புத் திட்டம்

இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்புத் திட்டம் இணையம் அடிப்படையிலான மொழிபெயர்ப்பு திட்டம் 2006 அக்டோபரில் மைய அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில் நுட்ப அமைச்சின் கீழ் தகவல் தொழில் நுட்பத் துறையின் நிதி நல்கையால் தொடங்கப்பட்டது. மலையாளம்-தமிழ்,

தமிழ்-மலையாளம் இயந்திர மொழிபெயர்பு தஞ்சாவூர் தமிழ்ப் பல்கலைத்திற்கு ஒப்படைக்கப்பட்டு முதன்மை ஆய்வாளர் என்ற நிலையில் இராசேந்திரனின் மேற்பார்வையில் நடைபெற்றது. இதன் முதல் நிலை 2006 அக்டோபர் 30-ஆம் நாள் தொடங்கப்பட்டு 2010 ஏப்பிரலில் முடிவுறது. இரண்டாவது நிலை IIITKM, நிருவனந்தபுரத்தில் மேற்கொள்ளப்பட்டது. இந்திய மொழியிலிருந்து மற்றொரு இந்திய மொழிக்கு இயந்திர மொழிபெயர்ப்பை உருவாக்கும் இணையம் அடிப்படையிலான மொழிபெயர்ப்புத் திட்டமாகும். இந்த ஒழுங்குமுறை இரு திசையில் செயல்படும் மற்றும் பொதுவான களத்தில் பின்வரும் மொழிகளுக்கிடையில் செயலாற்றியது.

தமிழ் - இந்தி
தெலுங்கு - இந்தி
மராத்தி - இந்தி
பெங்காளி - இந்தி
தமிழ் - தெலுங்கு
உருது - இந்தி
கன்னடா - இந்தி
பஞ்சாபி - இந்தி
மலையாளம் - தமிழ்

இந்திய மொழி பேசுபவர்கள் இந்த ஒழுங்குமுறையைப் பயன்படுத்துவார்கள். இந்த ஒழுங்குமுறையைக் கற்றல் நுட்பங்களைப் (Machine learning teachiques) பயன்படுத்தி மேலும் இயலும். இந்த ஒழுங்குமுறை இரண்டு குறிப்பிட்ட களங்களுக்கு இயந்திர மொழிபெயர்ப்பை செய்யும்.

1. சுற்றுலா மற்றும் புனிதப்பயணம் (Tourism and Pilgrimage)
2. நலவாழ்வு (Health)

பின்வரும் அடிப்படைக் கருவிகள் இந்திய மொழி மொழிப்பெயர்ப்பு ஒழுங்குமுறையின் பகுதியாக உருவாக்கப்பட்டுள்ளது.

1. சொல்வகைப்பாட்டு அடையாளப்படுத்தி (POS Tagger)
2. தொடர் பகுப்பான் (Chunker)

3. உருபனியல் பகுப்பாய்வி (Morph Analyser)

4. குறிப்பிட்ட இடம் சார்ந்த குழும் (Local word Grouper)

5. மாற்றமைவுப் பகுதி (Transfer Module)

6. இருமொழியப் பொருத்தம் (Billingual Mapping)

7. வாக்கிய நிலை உருவாக்கம் (Sentence Level Generation)

8. வட்டாரச் சொல் குழுமப் பிரிப்பான் (Local Word Group Splitter)

9. உருபனியல் உருவாக்கம் (Morphological Generation)

வாய்ப்பு (scope)

இந்த ஒழுங்குமுறை நிலைப்பேறு பெற்ற மொழியில் எழுதப்பட்ட இணைய பக்கங்கள் அல்லது கால இதழ்கள், நாளிதழ்கள் மீது வேலை செய்யும். சரியான தன்மை பயன்படுத்துவோர் திருப்தி (accuracy / user satisfaction) 85 % - 90 % விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிப்பெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்பட்டுள்ளது மற்றும் 80– 85 % திறமையுடன் பொதுப் பயன்பாடு இயந்திர மொழிபெயர்ப்பு உருவாக்கப்படுள்ளது.

சரியான தன்மை/ பயன்படுதுவோர் திருப்தி (Accuracy/user Satisfaction)

85-90 விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிபெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்படும். மற்றும் 80-85 விழுக்காடு திறமையுடன் பொதுப்பயன்பாடு இயந்திர மொழிபெயப்பு உருவாக்கப்படும்.

ஒழுங்குமுறைச் சிறப்பு (System Specification)

இந்திய மொழிபெயர்ப்புத் திட்டம் இணையத்துடன் தொடர்பு கொண்டு பயன்படுத்துவோருக்குத் தானியக்க மொழிபெயர்ப்பு சேவைஇயைத் தரவேண்டி ஒரு பரிமாறி (Server) இயங்கும். பின்வருவன மென்பொருள் சிறப்பீடுகளாகும்.

Plat form	:	Fedora Core 4 (Kernel – 2.6.11)
Web server	:	Apache – 2.0
Data base	:	GDBM – 1.8
Compiler	:	gcc 4.0

நெறிமுறை (Methodology)

இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறை பல இயற்கைமொழி ஆய்வு ஆய்வுக் குழுவினர் பங்களிப்பால் உருவாக்கப்பட்டுளது. இந்திய மொழி பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கம் பெரிய செயல்பாடு சிறிய செல்பாடுகளாப் பகுக்கப்பட்டுள்ளது. பங்களிக்கும் ஒவ்வொரு குழுவினரும் ஒன்றோ அதற்கு மேலோ வேளைகளை எடுத்துக்கொண்டு இணையாக வேலை செய்துள்ளனர். பெரும்பாலான பகுதிகள் மொழிச் சுதந்திரமான (language independent engine) இயந்திரமாகவும் மொழிச் சிறப்பான தரவாகவும் (language specific data) பகுக்கப்பட்டுள்ளது. இது போல் ஒரு மொழிக் குழுவினர் ஒரு குறிப்பிட்ட மொழிக்குத் தரவைத் தருவர். ஒரே இயந்திரம் பல மொழிகளுக்கு வேலை செய்யும்.

பகுதிகளாக அமைத்தல் (Modularity)

இந்திய மொழிபெயர்ப்பு ஒழுங்குமுறை பல பகுதிகளைக் கொண்டது. ஒவ்வொரு பகுதியும் ஒரு தர்க்கம்சார் வேலையைச் செய்யும், பெரும்பாலும் இந்த வேலை சிறியதாகும். இதன் காரணமாக ஏதாவது மாற்றம் ஏற்பட்டால் சிறிய பகுதியில் அதை எளிதாகச் செய்யவியலும்.

சக்தியின் தரமான அமைப்பு (Shakthi Standard Format (SSF)

எல்லாப் பாகங்களும் (modules)வடிவ அமைப்பு நிர்ணயிக்கப்பட்ட தரவின் ஒழுக்கில் செயல்படும். ஒரு பகுதி அதன் வேலையில் வெற்றி பெற்றால் அது வெளிPயீடு ஒழுங்குப் புதிய சின்னத்தையோ ஆய்வையோ சேர்க்கும். பெரும்பாலான பகுதிகளுக்குத் தனிப்பகுதிகளின் வட்டாரச் சிக்கல்களை உறுதியான கட்டுப்பாட்டில் வைக்க இந்த அணுகுமுறை உதவும்.

தோல்விகளை நேர்செய்தல் (Dealing with failure)

(SSF) வடிவமைப்பு பகுதியின் ஒவ்வொரு நிலையில் ஏற்படும் தோல்விகளை நேர்செய்ய வசதி செய்யும். எடுத்துக்காட்டாக அதே SSF -இன் மீது அதன் சின்னத்தின் மதிப்பை விளைவிக்கத் தவறினால் கீழ் ஒழுங்குப் பகுதிகள் (Down Stream Modules) தரவின் ஒழுங்கின் மீது தொடர்ந்து செயல்படும்.

தெளிவு (Transparency)

இந்தியமொழி இயந்திர மொழிப்பெயர்ப்பில் SSF வடிவமைப்பின் பயன்பாடு ஒவ்வொரு பகுதியின் உள்ளீட்டிற்கும் வெளிப்யீட்டிற்கும் முன் நிகழ்ந்திராதத் தெளப்வைப் பெற உதவுகிறது. ஒரு பகுதியின் உரை சார் வெளிப்யீடு மனிதப் பயன்பாட்டிற்கு மட்டுமல்லாமல் அதன் உள்ளீடாக தரவு ஒழுக்கில் அடுத்து வருகிற பகுதியிலும் பயன்படுத்தப்படும்.

டாஷ்போர்டு (Dashboard)

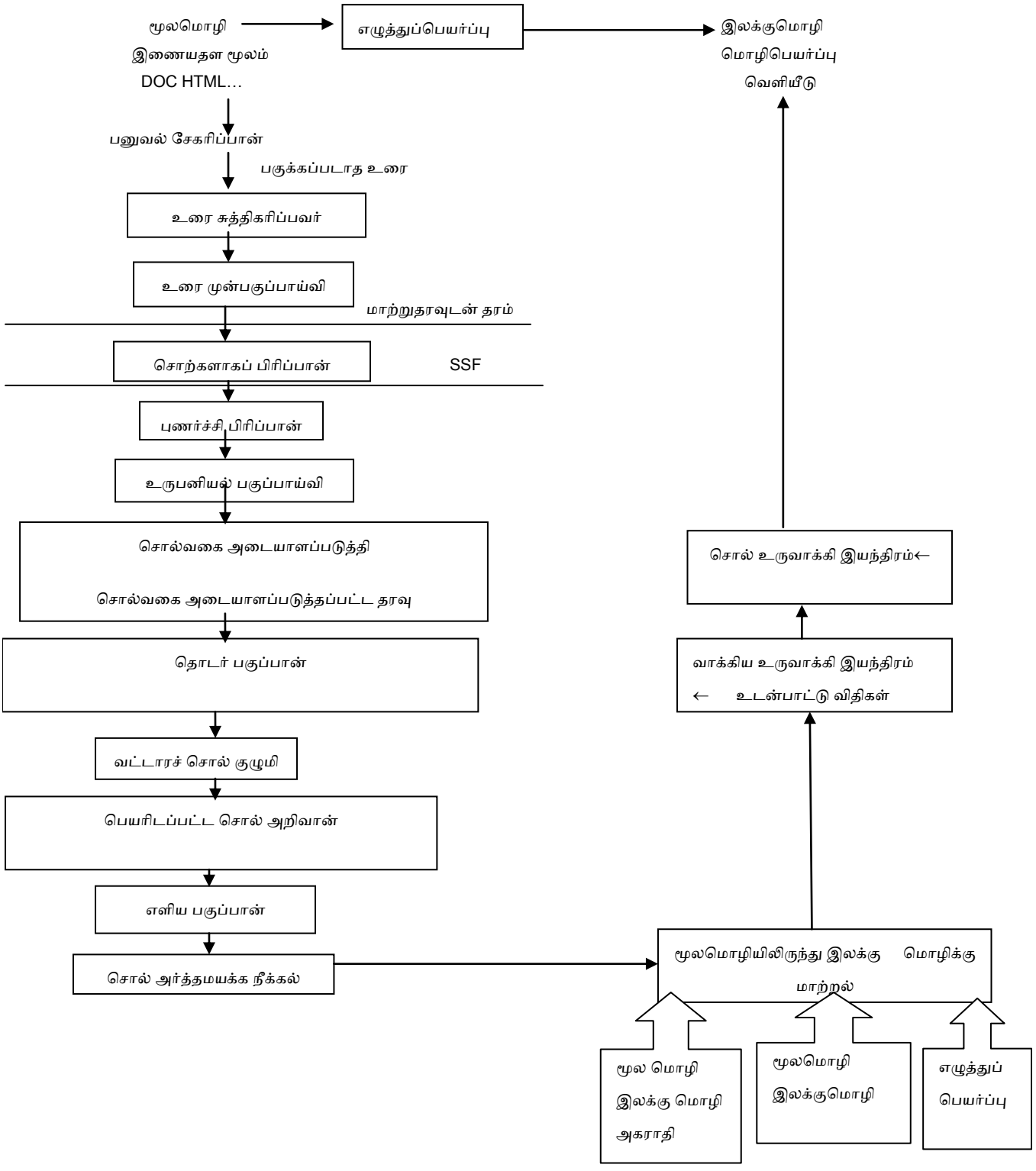
இந்திய மொழி மொழிபெயர்ப்பின் எல்லாப் பகுதிகளும் டாஷ்போர்டின் (dash board) உதவியால் இயங்குகிறது. இது குறிப்பிட்டபடி தரவு ஒழுக்கின் பைப்லைன்களை உருவமைப்பு செய்கிறது. இருப்பினும் பைப் லைன்களுக்குப் பதிலாக எல்லாப் பகுதிகளும் செயல்படும் பங்கிடப்பட்ட நினைவகத்தையோ பிளாக் போர்டையோ (Black Board) நிறுவுவதால் ஒங்குமுறையை விரைவுப்படுத்த இயலும். இருப்பினும் இது பங்கிடப்பட்ட நினைவாகத்தில் (அல்லது நினைவக உருப்படுத்தத்தில் SSF) இயங்குவதற்கு வேண்டி பகுதிகள் எழுதப்பட்டிருக்கவேண்டும் என்று அவசியப்படும்.

முழுவிளக்கம் (Overall Description)

இந்திய மொழியிலிருந்து இந்திய மொழிக்கான மொழிபெயர்ப்பு ஒழுங்குமுறை ஆய்தல் - மாற்றியமைத்தல் - உருவாக்குதல் என்ற மேற்கோள் வாய்ப்பாடு அடிப்படையில் செயல்படும். முதலாவது மூலமொழியின் ஆய்வு செய்யப்படும். பின்னர் இலக்குமொழிக்குச் சொற்றொகையும் ஆயப்பட்ட அமைப்பும் மாற்றம் செய்யப்படும். இறுதியாக இலக்கு மொழி உருவாக்கப்படும்.

இந்திய மொழிகள் ஒற்றுமையுள்ளதாக இருப்பதாலும் இலக்கண அமைப்புகளைப் பங்கீட்டுக் கொள்வதாலும் ஆழமற்ற பகுத்துக் குறித்தல் செய்யப்படும். மாற்றமைவு இலக்கணப் பகுதி எளிமையாக வைக்கப்படும். பொருண்மை களச் சிறப்புப் பெயரிடப்பட்ட சொற்களின் புரிந்து கொள்வான்களாலும் பொருத்தமான அகராதி போன்றவற்றாலும் கையாளப்படும். ஆழமற்ற பகுத்துக் குறித்தலானது, உருபனியல் ஆய்வு, சொல் வகைப்பாடு அடையாளப்படுத்தல் மற்றும் தொடர்பகுத்தல் என்பனவற்றை உள்ளடக்கும். முதலாவது (உருபனியல் பகுத்தாய்வு) விதி அடிப்படையிலானது, இரண்டாவது (சொல் வகைப்பாடு அடையாளப்படுத்தல்) புள்ளியியல் அடிப்படையிலானது, மூன்றாவது (தொடர்ப்பு) விதி அடிப்படை மற்றும் புள்ளியியல்

அடிப்படை இரண்டும் இணைந்தது. முழு ஆய்வு வேலையும் பல பாகங்களைப் பிரிக்கப்பட்டுள்ளது. ஒவ்வொரு பாகமும் சிறிய தர்க்க வேலையைச் செய்யும். இந்த ஒழுங்குமுறையின் முழு அமைப்பும் கீழே தரப்பட்டிருக்கும் படத்தில் காட்டப்பட்டுள்ளது.



இந்தியமொழியிலிருந்து இந்தியமொழிக்கு இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையின் படம்

ஒவ்வொரு முக்கியமான பகுதியும் கீழே விளக்கப்பட்டுள்ளன. ஒவ்வொரு பகுதியும் அதற்கு நேரே தரப்பட்டுள்ள முக்கிய வேலையைச் செய்கின்றன. இருப்பினும் முக்கிய வேலை தவிர அவை சில முன்னாய்வையும் (Preprocetsins) பின்னாய்வையும் (Post processing) செய்கின்றன. எடுத்துக்காட்டாகத் தொடர்பகுப்பான் பகுது தொடர் பகுதிகளைக் கண்டுபிடிக்கின்றன. இவ்வாறு தொடர்பகுப்பான் பகுது முக்கிய வேலையான தொடர் பகுத்தலைச் செய்கின்றது. இருப்பினும் தொடர் பகுதியின் தலைமையைக் கண்டுபிடிக்கும் பின்னாய்வைச் செய்கின்றது. மேலும் தலைமைப் பண்புக் கூறுகளைத் தொடர்பகுதியின் பண்புக்கூறுகளாக நகல் செய்கின்றது.

உருபனியல் பகுப்பாய்வி (Morphological Analyzer)

ஒரு சொல் தரப்படும் போது உருபனியல் ஆய்வி அதன் வேரையும் சொல்லின் இலக்கணப் பண்புக்கூறுகளையும் அடையாளங்கண்டு கொள்கிறது. திரிபு வள்ம் இல்லாத மொழிகளுக்கு எல்லாச் சொல் வடிவுகளையும் கொண்ட எளிய அகராதி நோக்கீடு போதுமானது. தமிழ், தெலுங்கு போன்ற திரிபு வளமுள்ள மொழிகளுக்கு அகராதியை உருவாக்குவது எளிதல்ல. இது பெரிய சேகரிப்பையும் உயர்ந்த செயல்பாட்டுக் கணிப்பையும் வேண்டும். இதற்கு நல்ல மாற்று வழி வேர்ச்சொற்களின் அகராதியை வைத்துக் கொண்டு உருபனியல் ஆய்வியைப் பயன்படுத்தி சொ வடிவை வேராகவும் இலக்கணப் பின்னுருபுகளாகவும் பிரிப்பதாகும்.

மேற்கோள் வாய்ப்பாடு அடிப்படையிலான அணுகுமுறைகள் (Paradigm based approaches) இந்திய மொழிகளுக்கு நல்ல வெற்றி மதிப்பீட்டைத் தருகிறது.

சொல்வகைப்பாடு அடையாளப்படுத்தி (POS Tagger)

சொல்வகை அடையாளப்படுத்தல் வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் ஒரு சொல் வகைப்பாட்டைத் தருவதாகும். வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் பெயர்கள், வினைகள், பெயரடைகள், வினையடைகள் என்று அடையாளம் தருவது ஒரு வாக்கியத்தில் ஒவ்வொரு உறுப்பு பின் ஒரு பங்களிப்பை ஆய்வதற்கு உதவும். சொல்வகைப்பாட்டு அடையாளப் படுத்துவதற்கு விதி அடிப்படை, புள்ளியியல் அடிப்படை போன்ற பல அணுகுமுறைகள் இருக்கின்றது. இங்கு புள்ளியல் அணுகுமுறை பின்பற்றப்படுகின்றது.

தொடர் பகுப்பான் (Chunker)

தொடர்பகுத்தல் (Chunking) ஒரு வாக்கியத்தில் உள்ள எளிய பெயர்த்தொடர்கள், வினைக்குழுமங்கள், பெயரடை தொடர் மற்றும் வினையடை தொடர்களை அடையாளம் காண்பது ஆகும். இது தொடர் பகுது மற்றும் புலக்குறிப்பு இவற்றின் எல்லையை அடையாளங்காண்பதை உட்படுத்தும்.

குறிப்பிட்ட இடம்சார் குழுமி (Local Word Grouper)

வட்டாரச் சொல் குழுமி விபக்தி கணினியக்கத்தின் (Vibhakti Computation) தொழில்நுட்ப வேலையைச் செய்யும் தனிச்சொற்களாக வரும் துணை வினைகள் தனி பொருளடக்கச் சொற்களாகக் கருதப்படுவதில்லை. ஏனென்றால் அவை முதன்மை வினையின் தலைமையாக இருக்கும் தொடர்பகுதியின் இலக்கணப் பண்புக் கூறுகளைச் சிறப்பீடு செய்யும். இது இந்தியில் தனிச் சொற்களாக வருகின்ற பெயர்விபக்தி என்று அழைக்கப்பெறிகின்ற பின்னுருபுகளுக்கும் உண்மையாகும். வட்டாரச் சொல் குழுமி பொருத்தமான கலவை நிலை மாறும் கூட்டுநிலை சொல் குழும ஆக்க விதிகளைத் திறமையாகக் கையாளும். முக்கியமான வேலை வட்டாரத் தகவல் அடிப்படையில் செயல்பாட்டுச் சொற்களைப் (Functional Words) பொருளடக்கச் சொற்களுடன் (Content Words) குழுமுவதாகும். இந்த வேலையை நிறைவேற்ற பயன்படுத்தப்படும் ஒரு செயல்முறை கிரியரூப்பா அட்டவணைகளைப் (Kiya rupa Charts) பயன்படுத்துவதாகும். இந்த அட்டவணைகள் ஒரு தனி செயல்பாட்டைக் குறிக்கும் தொடர்ச்சியான வினைகளிலிருந்து உருவாக்க வேண்டிய குழுமங்களைச் சிறப்பீடு செய்யும்.

மாற்றம் பகுதி (Transfer Module)

மொழியாய்வில் வாக்கியங்கள் அவற்றின் தொடரமைப்பை அடையாளம் காண்பதற்காகப் பகுத்துக் குறிக்கப்பெறும். இந்திய மொழிகளுக்குள் வேறுபாடுகளை விட ஒற்றுமைகள்தான் அதிகம். எடுத்துக்காட்டாகத் தமிழ், இந்தி மொழியினை முழு பகுத்துக் குறித்தலை வேண்டாது. இத்திட்டத்தில் இயந்திர மொழிபெயர்ப்பு முழு வாக்கியப் பகுத்துக் குறிப்பான் இல்லாது நிறைவேற்றப்படும். மூலமொழியின் அமைப்பிற்கு இலக்குமொழி அமைப்பில் நிகரான அமைப்பு இல்லாவிட்டால் அமைப்பு சார் மாற்றம் தேவைப்படும். ஒரு பகுதி பகுத்துக் குறித்தல் அல்லது ஆழமற்ற பகுத்துக் குறித்தல் மாற்றாத்திற்கு உட்படுத்தப்படும் வாக்கியங்களின் குறிப்பிட்ட உறுப்புகளை அடையாளம் கண்டால் போதுமானது.

மொழிகளுக்கிடையில் உள்ள தொடரியல் வேறுபாடு கலவை வாக்கிய கட்டுமானங்களில், குறிப்பாக எச்சத்தொடர்களில் காணப்படும். இது மாற்றமைவு இலக்கணத்தால் இணைக்கப்படும். இந்தப் பகுது எழுத்துருவாக்கத்தின் வேலையையும் உட்படுத்தும். இந்திய மொழிகளுக்குள் (உருது மொழியை உள்ளடக்கி) எழுத்துருவாக்கத்தை நிறைவேற்றும் ஒரு பகுதி உருவாக்கப்பட வேண்டும். எழுத்துருவாக்கம் சொல்லையோ சொற்களையோ படிப்பவரின் எழுத்துக்கு மாற்றுகிறது. எடுத்துக்காட்டாக ஒருவருக்கு இந்து தெரிந்திருந்தால் தேவநகரியில் வங்காள உரையைப் படிக்க இயலும் என்றால் அதன் பொருளின் சில பாகங்களை அவரால் புரிந்து

கொள்ள முடியும். மொழிபெயர்ப்பு ஒரு சொல்லுக்கோ ஒரு தொடர் பகுதிக்கோ தோற்றுப்போனால் எழுத்துருவாக்கம் படிப்பவரை படிக்கவும் புரிந்து கொள்ளவும் அனுமதிக்கும். இந்திய மொழிகள் பல எண்ணிக்கையிலான சொற்களைப் பங்கிட்டுக் கொண்டுள்ளன. எனவே படிப்பவருக்கு ஏற்றவாறு எழுத்துருமாற்றம் செய்தால் அவரால் மூலமொழியில் உள்ள உரையிலிருந்து சிலவற்றைப் புரிந்து கொள்ள இயலும்.

இருமொழிய அகராதி பொருத்தம் (Bilingual Mapping)

உருபனியல் பகுப்பாய்வால் அடையாளம் காணப்பட்ட வேர்ச்சொல் இலக்குமொழி நிகரன்களுக்கு வேண்டி இருமொழி அகராதியில் பார்க்கப்படும். இந்த இருமொழி அகராதி இலக்கு மொழியின் வேர்ச்சொல் நிகரன்களையும் அதன் வகைப்பாடு (பெயர், வினை போன்றவை) மேற்கோள் வாய்ப்பாடு மற்றும் பிற தேவையான தகவல்களைக் கொண்டிருக்கும். இந்த நிலை மூலமொழி இலக்கண மின்னுருபுகளுக்கு இலக்கு மொழி நிகரன்களை அடையாளம் காண அகராதியைப் பயன்படுத்துகிறது.

வாக்கியநிலை உருவாக்கம் (Sentence Level Generation)

இந்தச் செயல்பாடு தேவையானால் தொடர்பான்களை முதன்மையாக நிரல் மாற்றியமைக்கும். சில வாக்கியநிலை நோக்குகளும் இந்தப் பகுதியால் கையாளப்படும். எடுத்துக்காட்டாக பெயர், வினை உடன்பாடு போன்றவை இங்கு கையாளப்படும். இந்த மட்டம் தொடர்பகுதி மட்ட ஆய்வாகும். வட்டாரச்சொல் வகுப்புகளைக் கொண்ட தொடர் பகுதிகளுக்கு பொருத்தமான சொற்கள் (அப்படியிருந்தால்) உருவாக்கப்படும். முதன்மை வினை, துணை வினை உடன்பாடு போன்ற உள் தொடர் பகுதி உடன்பாடு இங்கு காணப்படும்.

குறிப்பிட்ட இடம்சார் குழு பிரிப்பான் (Local Word Group Splitter)

இந்தப் பகுதி ஒரு சொல்லுக்கு விபக்தி மற்றும் பிற தகவல்களைக் கணிக்கும் மற்றும் அவற்றைத் தனிச்சொற்களாகத் திருப்பி அனுப்பும்.

உருபனியல் உருவாக்கம் (Morphological generation)

ஒரு குறிப்பிட்ட இலக்கண வகைக்காக ஒரு சொல்லை உருவாக்க வேர்ச்சொல்லுடன் பொருத்தமான பின்னொட்டுகள் இணைக்கப்பட வேண்டும். உருவாக்கி மொழியின் உருபனியலுக்குத் தக்கவாறு வேர்ச்சொல்லை திரிபுற செய்யும் மற்றும் இலக்குமொழி சொல் வடிவை வெளியிடும். இவ்வாறு உருவாக்கப்பட்ட சொற்கள் முழு இலக்குமொழி வாக்கியத்தை உருவாக்க இணைக்கப்படும்.

மதிப்பீடு செய்தல் (Evaluation)

மதிப்பீடு செய்யும் இயக்க நுட்பம் தனி வாக்கியங்களுக்கும் உரைகளுக்கும் வேண்டி இறுதி பயன்பாட்டாளர்களுக்காக நிறுவப்படும். இது புரிந்து கொள்ளுதல் மற்றும் மொழிபெயர்ப்பு தரம் இவற்றிற்காக மதிப்பீடு செய்யப்பட்டுள்ளது.

இரண்டாவது நிலையில் மதிப்பீடு ஒழுங்குமுறையின் வேகத்திற்காகவும் பயன்பாட்டிற்காகவும் மேம்படுத்தப்பட்டு மதிப்பீடு செய்யப்பட்டுள்ளது.

10.7.2.6.. கூகுளின் மொழிபெயர்ப்பி

கூகுள் ஆங்கிலத்திலிருந்து தமிழுக்கும் தமிழிலிருந்து ஆங்கிலத்திற்கும் மொழிபெயர்ப்பதற்கு ஒரு மொழிபெயர்ப்பியை இணைய தள ஆன்லினில் தருகின்றது. ஆங்கில-தமிழ் மொழிபெயர்ப்பி சிறப்பாக மொழிபெயர்ப்புச் செய்கின்றது. இது புள்ளியியல் நெறிமுறை அடிப்படையில் இணையான ஆங்கில-தமிழ் விரிதரவுகளைப் பயன்படுத்தி மொழிபெயர்ப்பு செய்வதாய்த் தெரிகின்றது.

10.8. சுருக்க உரை

இயந்திர மொழிபெயர்ப்பு இன்றைய காலகட்டத்தில் முக்கியமான ஒரு நடைமுறை செயல்பாடாக மாறும் நிலையை நோக்கி வளர்ந்து கொண்டிருக்கிறது. இணைய தளத்தில் ஏராளமான தகவல்கள் தமிழிலும் பிறமொழிகளிலும் இருக்கின்றன. குறிப்பாக ஆங்கிலத்தில் பல முக்கியமான தகவல்கள் இணைய் தளத்தில் வலம் வருகின்றன. இத்தகவல்களை தமிழ் மட்டும் அறிந்த நம்மால் புரிந்து கொள்ள இயலாது. இதற்கு வேண்டி உடனடி மொழிபெயர்ப்பு தேவை. இணைய தளத்தில் இருக்கும் ஆங்கிலத் தகவல்களை உடனடியாக இயந்திரத்தினாலேயே தமிழில் மொழிபெயர்ப்பு செய்து கிடைக்கப் பெற்மானால் அது மிகுந்த வரபிரசாதமாகும். அத்தகைய திசை நோக்கித்தான் தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகிறது. தமிழிலிருந்து ஆங்கிலத்திற்கும் ஆங்கிலத்திலிருந்து தமிழுக்கும் தமிழிலிருந்து பிற மொழிகளுக்கும் தமிழ்மொழிக்கும் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகின்றன. தமிழ்ப் பல்கலைக்கழகம், அண்ணா பல்கலைக்கழகம், சென்னைப் பல்கலைக்கழகம், AUKBC ஆய்வு நிறுவனம், அமிர்தாப் பல்கலைக்கழகம் என்பன தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் ஈடுபட்டுள்ளன. இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சி (Technological Development Indian Languages) என்ற திட்டத்தின் கீழ் இந்திய மைய அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில்நுட்ப அமைச்சின் (Ministry Communication and Information Technology) கீழ்வரும் தகவல் தொழில்நுட்ப துறையால் பொருளாதார உதவிப்பெற்று பல நிறுவனங்களும் பல்கலைக்கழகங்களும் ஆங்கில

மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு இந்திய-இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு என்பனவற்றில் முழுமூச்சுடன் ஈடுபட்டுள்ளன. இம்முயற்சிகள் யாவும் இணையத் தளத்திலிருந்து மொழிபெயர்ப்பை இலக்காகக் கொண்டு அமைகிறது. இந்திய மொழிகளின் மைய நிறுவனம் இந்திய மொழிகளின் பேச்சு மற்றும் எழுத்து விரிதரவுகளை திரட்டும் திட்டத்தை பெற்று செயல்படுத்தி வருகிறது.

முற்றிலும் தானியக்க மொழியியல் பெயர்ப்பு இன்றைய காலகட்டத்தில் இயலாத வகையால் கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பை ஊக்குவிக்கிறது. பொதுவான மொழிக்கு அதாவது பொதுவான மொழி உரைக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவது கடினமாகையால் ஒரு குறிப்பிட்ட பொருண்மைகளை வைத்துக் கொண்டு இயந்திர மொழிபெயர்ப்புச் செய்யும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. ஆங்கில-இந்திய மொழிகள், இந்திய மொழிகள்-இந்திய மொழிகள் மொழிபெயர்ப்பு திட்டத்தில் சுற்றுலா மற்றும் உடல் நலம் (Health) என்பன பொருண்மை களங்களாக ஏற்கப்பட்டு மொழிபெயர்ப்பு செய்யும் முயற்சி மேற்கொள்ளப்பட்டு வருகின்றன.

இயந்திர மொழிபெயர்ப்பு என்பது காலத்தின் கட்டாயமாகும். தமிழுக்கான தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் நாம் ஈடுபட்டே ஆக வேண்டும். இல்லாவிட்டால் இருமொழிய (Bilingual) (ஆங்கிலம், தமிழ்) நிலையிலிருந்து ஒரு மொழிய நிலைக்கு (Mondingual)மாறிவருகின்ற நம் மாணவ சமுதாயத்தைக் காப்பாற்ற இயலாது போய்விடும். இணைய தளத்தில் கிடைக்கும் அறிவு மூலங்களைத் தமிழர்களாய நாம் பெறாது போய் விடுவோம். இதுபோல தமிழில் கிடைக்கும் அரிய தகவல்கள் பிற மொழியினவர்களுக்குப் போய்ச் சேராது. தகவல் பன்மடங்காகப் பெருகிவரும் இக்கால கட்டத்தில் தகவல் பரிமாற்றத்திற்கு இயந்திர மொழிபெயர்ப்பை விடையாகும்

இயல் 11
மொழிக் கருவிகள் உருவாக்குதல்

11.1.அறிமுகம்

இவ்வியலில் இதுவரை கூறப்பாத மொழிக்கருவிகள் உருவாக்குவது பற்றிச் சுருக்கமாக விளக்கப்படும். மொழிக் கருவிகளைப் பின்வருமாறு பிரித்து அறியலாம்:

1. ஒலிதொடர்பான பயன்பாட்டுக் கருவிகள்
2. எழுத்து தொடர்பான பயன்பாட்டுக்கருவிகள்
3. சொல்தொடர்பான பயன்பாட்டுக் கருவிகள்
4. மொழிபெயர்ப்புக் கருவிகள்
5. கற்றல் கற்பித்தல் கருவிகள்

சொற்கள் தொடர்பான பயன்பாட்டுக் கருவிகளைத் உருவாக்குவதற்குக் கணிப்பொறியைப் பயன்படுத்தலாம். பின்வருவன இதில் அடங்கும்.

1. சொற்களை அகரவரிசைப்படுத்தல்
2. சொல்லடைவு தயாரித்தல்
3. சொல்லாட்சி அகராதி மற்றும் பொருளடைவு தயாரித்தல்
4. அகராதி தயாரித்தல்
5. சொற்களஞ்சியம் தயாரித்தல்
6. எழுத்து மற்றும் இலக்கணப் பிழை திருத்தி தயாரித்தல்
7. சொல் ஆய்வி தயாரித்தல்

11.2.ஒலிபெயர்ப்பி உருவாக்குதல்

கணிப்பொறியில் எழுதப்பட்ட பனுவலை ஒலிப்பு வடிவில் வரும்படி செய்தலும் முக்கியமான ஒரு ஆய்வாகும். இத்தகைய ஒலிபெயர்ப்பிகளைப் பனுவலிலிருந்து பேச்சாகப் புரிந்துகொள்ள உதவும் கருவியை Text to Speech Recognizer என்பர். தமிழில் இதற்கான முயற்சிகள் தொடங்கப்பட்டு ஓரளவுக்கு வெற்றியும் கிட்டியுள்ளது.

11.2.எழுத்துப் பெயர்ப்ப்பி உருவாக்குதல்

ஒரு மொழியின் எழுத்துக்களால் எழுதப்பட்ட பிற மொழிகளின் எழுத்துக்களைப் பயன்படுத்திப் படிக்கத்தக்க முறையில் பெயர்ப்பு செய்யும் வசதிமுறைகள் கணிப்பொறியில்

வந்துவிட்டன. இதனால் தமிழ் எழுத்தில் எழுதப்பட்ட பனுவல்களை ரோமன் எழுத்தில் படிக்கவோ ரோமன் எழுத்தில் எழுதப்பட்ட பனுவல்களைத் தமிழ் எழுத்தில் படிக்கவோ இயலும்.

11.4.எழுத்துணரி உருவாக்குதல்

அச்சிடப்பட்ட பனுவல்களை அப்படியே கணினிக்கு எழுத்துக்களாகத் தானாகவே உள்ளீடு செய்ய இயலும் கருவியைப் எழுத்துணரி (பார்வை எழுத்தறிவான்) (Optical Character Recognitior) என்பர். இக்கருவி தட்டச்சு முறைப்படி சாவிப்பலகை வழி பனுவல்களை உள்ளீடு செய்ய ஆகும் கால விரையத்தை இல்லாமல் செய்கின்றது. ஆங்கிலத்திற்கு இத்தகைய கருவி உருவாக்கப்பட்டுப் பயன்படுத்தப்படுகின்றது. ஆனால் தமிழுக்கு அதன் எழுத்துக்களின் கலவைத் தன்மை காரணமாக இத்தகைய கருவியை உருவாக்குவது கடினமான செயலாகும். இருப்பினும் தமிழுக்கு இத்தகைய பார்வை எழுத்தறிவான் உருவாக்கும் முயற்சிகள் எடுக்கப்பட்டு ஓரளவுக்கு வெற்றியும் கிட்டியுள்ளது.

11.5.சொற்களை அகரவரிசைப் படுத்தல்

தமிழ்மொழியின் அகரவரிசை தொடர்பான செய்திகளையும் வழியமைப்பு ஆணைகளையும் (Program Instructions) கணிப்பொறிக்குத் தந்தால் கணிப்பொறி மிகக் குறைந்த நேரத்தில் வேண்டிய சொற்களை அகர வரிசைப் படுத்தித் தந்துவிடும். ஆங்கிலத்திற்கு இது மிகச் சாதாரண ஒன்று. தமிழைப் பொறுத்தவரையில் பலவிதமான எழுத்துருக்களும் (Fonts) அதற்குப் பின்னணியாய் அமையும் வேறுபட்ட குறியன்களும் (Codes) குறியன் எண்களும் (Code Numbers) பயன்படுத்தப்படுவதால் அகரவரிசைப் படுத்தல் எளிய செயலாக அமையவில்லை. இருப்பினும் வேறுபட்ட எழுத்துருக்களை வழியமைப்புகள் மூலம் இணைத்து அகரவரிசைப்படுத்தும் மென்பொருள்கள் கண்டுபிடிக்கப்பட்டுப் பயன்பாட்டில் உள்ளன.

11.6.தமிழ்ச் சொல்லாய்வி/சொல்லாளர் உருவக்குதல்

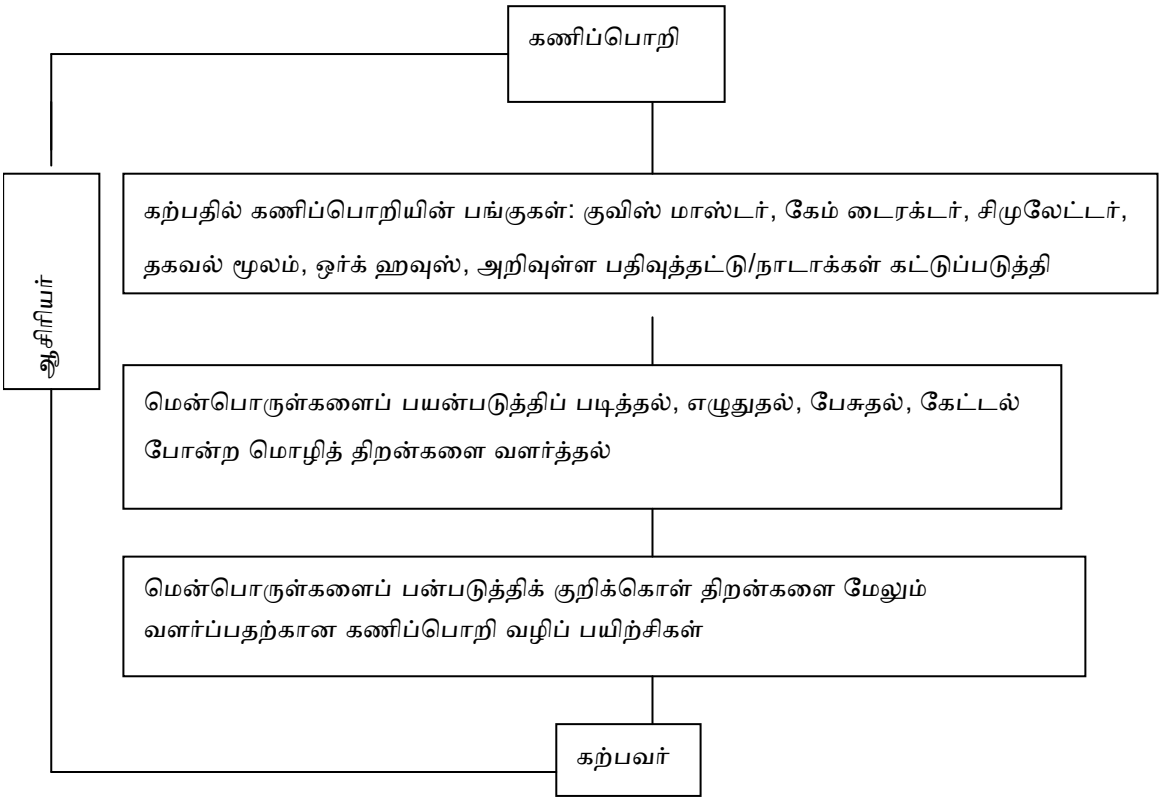
தமிழுக்கான சொல்லாய்விகள் பல இன்று வெளி வந்துள்ளன. அவை தமிழ்ப்பனுவல்களை உள்ளீடு செய்து வடிவமைத்து வெளியீடாகப் பெறும் நுட்பங்களைக் கொண்டுள்ளன. Word, Word Star போன்ற மென்பொருள்களில் கிடைக்கும் எல்லா வசதிகளும் இவற்றிலும் உள்ளன. சொற்களஞ்சியம், எழுத்து மற்றும் இலக்கணப் பிழைதிருத்தி என்பன மேலும் மேம்படுத்தப்படவேண்டும். பல தமிழ்ச் சொல் சொல்லாய்விகளில் இவ்வசதிகள் இல்லை.

11.7.மொழிகற்றல்/கற்பித்தல் கருவிகள் உருவாக்கல்

மொழியைக் கற்பதற்கும் கற்பிப்பதற்கும் கணிப்பொறியைப் பயன்படுத்தலாம். வெறும் எண், எழுத்து இவற்றை மட்டும் கையாண்டு வந்த கணிப்பொறி எண், எழுத்து, படம் (graphics), ஒலி (audio), ஒளி (video) அசையும் உருவங்கள் (animation) எனப்பலவற்றை ஏற்றுக்கொண்டுள்ளது. இப்படி வேறுபட்ட பன்முகத் தொடர்புகள் கொண்ட ஒருங்கிணைப்பு

மல்டிமீடியா (Multimedia) எனப்படும். மல்டிமீடியாவைப் பயன்படுத்தித் தமிழ்க்கற்றலையும் கற்பித்தலையும் ஆர்வத்தைத் தூண்டும் செயலாக மாற்றலாம்.

கணிப்பொறி அடிப்படையிலான கணிப்புச் செயல்களைக் கல்வி மற்றும் உளவியல் அடிப்படையிலான கற்றலுக்கும் கற்பித்தலுக்கும் பயன்படும் படி செய்வதுதான் கணிப்பொறி மூலம் கற்றல்/கற்பித்தல் மென்பொருள்களைத் தயாரித்தலின் நோக்கமாகும். இங்குக் கணிப்பொறி, கற்றல் அடிப்படையிலான மொழிச்செயல்பாடுகளைச் செய்வதற்குப் பல பங்குகள் வகிக்கும். எடுத்துக்காட்டாக குவிஸ்மாஸ்டர், கேம் டைரக்டர், சிமுலேட்டர், ஒர்க் ஹவுஸ், தகவல் மூலம் கட்டுப்படுத்தி போன்ற பல பங்குகளைக் கணிப்பொறி வகிக்கும். பின்வரும் வரைபடம் கணிப்பொறியின் உதவியுடன் கூடிய கற்றல்/கற்பித்தல் செயற்பாட்டைத் தெளிவுபடுத்தும்.



மேற்கண்டவாறு கணிப்பொறியைப் பயன்படுத்தி மொழித்தொடர்பான செயற்பாடுகளைச் செய்ய மொழியியலாரும் கணிப்பொறியியலாரும் இணைந்து செயலாற்ற வேண்டும். ஒருவருடைய செயற்பாட்டை மற்றவர் புரிந்துகொண்டு செயலாற்றினால் மொழித்தொடர்பான கணிப்பொறிக் கருவிகளை, மென்பொருள்களை எளிதில் உருவாக்கிப் பயன்படுத்தலாம்.

தமிழில் கணிப்பொறி வழி மொழி கற்றல், கற்பித்தல் மக்கள் முன்னேற்றத்திற்கு கணிப்பொறி பல வகையலி உதவுகிறது. தொடக்கக் காலங்களில் கணிப்பொறி நல்ல கணக்கீட்டு கருவியாக (calculating device) பயன்பட்டது. அன்றாட பணிகளுக்கு அல்லது பணிகளிலும் வழக்கமாக செய்ய நேரிடும் பணிகளிலும் கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறியின் வேகம், துல்லியம், செய்திகளை மீளத்தரும் வேகம் மற்றும் நினைவகத் திறன் போன்றவை அதிகரித்தன. ஆனால் அதற்கு எதிர் மாறாக விலை குறைந்தது. அதன் விளைவாகக் கணிப்பொறி பயன்பாடு அதிகரித்தது. ஆக்கப் பூர்வமான (creative) பணிகளுக்கும் கணிப்பொறிகள் பயன்படும் வண்ணம் உருவாக்கப்படுகின்றன. குறிப்பாகக் கல்விக் கூடங்களில் பயிற்றுவிப்பதற்கு உதவும் கருவியாக கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன. கணிப்பொறி வழி பாடங்களை தயாரித்து உரிய திட்டங்களை செயல்படுத்தி மாணவர்களுக்கு கற்பிக்கலாம். பாடங்களைக் கற்பதற்கும் கற்பிப்பதற்கும் கணிப்பொறி உதவுகிறது.

11.7.1. கற்றல், கற்பித்தலில் கணிப்பொறி

அடிப்படையில் கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. தேவையான பணிக்கான வழியமைப்பு அல்லது மென்பொருள் கொண்டு அது இயக்கப்படுகிறது. இருப்பினும் செய்திகளை சேகரித்தல், ஆராய்தல், வெளியிடுதல் மற்றும் பரப்புதல் போன்ற செயல்பாடுகளில் மிக மிக அதிவேகமுடையது. மேலும் கணிப்பொறியில் உரைச் செய்திகள், வரைபடங்கள் புகைப்படங்கள் ஒலிச்செய்திகள் மற்றும் சலனப்படங்கள் (animation) போன்ற பலவகையான செய்திகளை உள்ளீடு செய்து ஆராய்ந்து முடிவு பெறுவதற்கான வசதிகள் உள்ளன. எல்லா வகை செய்திகளையும் கலந்த கலவையாகவும் செய்திகள் அமையப்பெறலாம். இது போன்று எல்லா வகை செய்திகளையும கலந்த கலவையை பல்லூடகம் (multimedia) என்கிறோம்.

மற்றொரு பக்கத்தில் பயிற்றுவிப்போர் அல்லது ஆசிரியர்களுக்கு இருக்க வேண்டிய அடிப்படை இயல்புகள் மூன்று அவையாவன:

1. செய்திகளைத் திரட்டுதல்
2. செய்திகளை ஆராய்தல்
3. பயில்வோருக்கு ஏற்றவாறு செய்திகளை எடுத்துரைத்தல்

இத்தகைய குணங்களைக் கணிப்பொறி பெற்றிருப்பதால் கற்பித்தலுக்கும் கற்பதற்கும் சிறந்த சாதனமாகப் பயன்படுத்தப்படுகிறது. கணிப்பொறியால் சொந்தமாகப் பாடங்களை உருவாக்கவோ பாடங்களை நல்ல முறையில் வடிவமைப்பதற்கோ இயலாது. இவ்வாறான கணிப்பொறியை ஒரு ஆசிரியரைப் போன்று செயல்பட வைப்பதற்கு மென்பொருள் (soft ware)

உருவாக்கப்படுகின்றன. ஒரு பாடத்தைப் பயிற்றுவிப்பதற்காக உருவாக்கப்படும் மென்பொருள் பொதுவாக பாட மென்பொருள் (course ware) எனப்படுகிறது.

கணிப்பொறி வழி கற்பித்தலை ஆங்கிலத்தில் CAI (Computer Assisted/ Aided Instruction) என்று கூறுவர். இங்கு assisted அல்லது aided என்ற சொற்கல் கற்பித்தலுக்கு கணிப்பொறி உதவும் என்ற கருத்தை வலியுறுத்தும். அதுபோல் கணிப்பொறி மூலம் பயிலும் முறையை ஆங்கிலத்தில் CAL (Computer Aided Learning) என்று கூறுவர். CAL மற்றும் CAI போன்ற சொற்களுக்கு நிகராக பல தொடர்கள் பயன்படுத்தப் படுகின்றன. CMI (Computer Managed Instruction), CML (Computer Managed Learning), CDL (Computer Directed Learning) என்பன சிலவாகும்.

11.7.2. கணிப்பொறி வழி கற்றல்-கற்பித்தலின் சிறப்பியல்புகள்

கணிப்பொறி வழி கற்றல், கற்பித்தலில் எண்ணற்ற பயன்களும், சிறப்பியல்புகளும் உள்ளன. பொதுவாக இச்சிறப்பியல்புகளை மூன்று வகைகளாகப் பிரிக்கலாம். அவையாவன,

1. கணிப்பொறியில் இயல்பாக இருக்கின்ற சிறப்பு குணங்கள்.
2. பயில்வோருக்கு கிடைக்கின்ற பயன்கள்.
3. பயிற்றுவிப்போருக்கு உதவும் சிறப்புகள்.

கணிப்பொறியின் செயல், வேகம், துல்லியம், சோர்வின்மை, சகிப்புத்தன்மை, நெகிழ்வுத் தன்மை, ஒரே நேரத்தில் பல செயல்களை புரியம் திறமை போன்றவை கணிப்பொறியில் இயல்பாக அமைந்துள்ள பண்பு நலன்களாகும். கணிப்பொறிகள் மனிதனைப் போல் சோர்வடைவதில்லை. எனவே எப்பொழுது வேண்டுமானாலும் எவ்வளவு நேரம் வேண்டுமென்றாலும் தொடர்நது பயன்படுத்தலாம். எனவே, கணிப்பொறிகள் கல்வி கற்றல் கற்பித்தலில் பயன்படுத்தப் பட்டு வருகின்றன. மேலும் வேறு எந்த ஒரு கருவியிலும் இல்லாத பல்லூடக சாதன அமைப்புக் கொண்டது. எனவே உரைநடைச் செய்திகள் மட்டுமல்லாது வரைப்படங்கள், புகைப்படங்கள் மற்றும் ஒலிச் செய்திகள் போன்றவைகளை ஆராய்வதற்கும், கையாளவதற்கும் எண்ணற்ற வசதிகளைப் பெற்றுள்ளது. இதனால் கல்வி கற்பிப்பதற்கும் கற்பதற்கும் தேவையான அனைத்து அடிப்படை வசதிகளையும் பெற்றுள்ளது.

பட வில்லைகள் (slides) புகைப்படங்கள் மற்றம் கல்வித் தொடர்பான சினிமா படங்கள் கற்றலுக்கும், கற்பித்தலுக்கும் சிறந்த சாதனங்களாக பயன்படுத்தப்பட்டு வருகின்றன. இவைகளை கருவியாக பயன்படுத்திக் கொள்ள இயலுமெனினும் அந்த நிகழ்ச்சிகளுடன் உரையாடுவதற்கோ, கற்போரின் தன்மைக்கேற்ப கட்டுப்படுத்தவோ இயலாது. அந்நிகழ்ச்சிகளை

தயாரித்த வரிசை கிரம முறையில் மட்டுமே பார்க்கலாம். கற்போரின் மனநிலைக்கு ஏற்ப தேவையான செய்திகளை கூடுதலாகவோ அல்லது குறைவாக பெறவோ இயலாது. ஆனால் கணிப்பொறி வழி இத்தகைய கல்வி நிகழ்ச்சிகள் அல்லது பாடங்களைப் பதிவு செய்யும் பொழுது அதனுடன் உரையாட இயலும். தேவையான செய்திகளை தேவையான வண்ணம் கேட்டுப் பெறலாம். இது கணிப்பொறி கருவியின் மிகப் பெரிய சிறப்பம்சமாகும்.

கணிப்பொறியை எந்நேரமும் எப்பொழுதும் பயன்படுத்தலாம் என்பதால் பயில்வோர் விருப்பமான நேரத்தில் விருப்பமான பாடத்தைப் பயிலலாம். கணிப்பொறி வழி பயில்வோரின் சுரந்திரம் தேவை. இங்கு பயில்வோர் தனித் தனியாக பயில வாய்ப்புகள் உண்டு. 1.பயில்வோரின் ஈடுபாடு, 2. கற்கும் திறன், 3. கற்கும் வேகம் ஆகிய தன்மைகளைப் பொறுத்து பாடங்களைக் கற்றுக் கொள்ளலாம். சாதாரண வகுப்புகளில் இத்தகைய சாத்தியங்கள் இல்லை. பாடங்களை விரைவாக கற்றுக் கொ;ளபவர்கள் மெதுவாக கற்கு கொள்பவரின் வேகத்திற்கு ஒத்துப்போக வேண்டும். இதனால் வேகமாக கற்றுக் கொள்பவரின் நேரம் வீணாகி போகும். வகுப்பில் கூச்ச சுபாவத்தால் ஐயப்பாடுகளை களைவதற்கான கேள்விகளை கேட்பதற்கு மாணவர்கள் தயங்குவர். ஆனால் கணிப்பொறி வழி கற்கின்ற பொழுது மனநிலையைப் பொருத்து ஒரு பாடத்தில் அதிக செய்தியையோ, பொருளையோ, முக்கிய செய்தியையோ படித்து அறிவதற்கான பரிமாணங்களை அல்லது சிறப்புகளைப் பயில்வோரே நிர்மானித்துக் கொள்ளலாம். மேற்கூறியவைகள் எல்லாம் கணிப்பொறி வழி பயில்வோருக்கு உள்ள சிறப்புத் தன்மைகள் ஆகும்.

கணிப்பொறி வழி கற்பிப்பதற்கான பாடங்களை உருவாக்கும் ஆசிரியர்கள் மாணவர்களின் கல்வி நிலைக்கு ஏற்ப, கற்கும் திறனுக்கு ஏற்ப வடிவமைக்க வேண்டடும் என்பதில்;லை. ஒரு பொதுவான பாடங்களைத் தயாரித்து பயில்வோரின் திறமைக்கு ஏற்ப அவைகளை உருவாக்கித் தர இயலும். உரைநடைச் செய்திகள், விளக்கப் படங்கள் அனைத்தும் இதில் உள்ளதால் பாடக் கருத்துகளை எளிதாக புரிய வைக்க இயலும். ஒருமுறை பாடத் திட்டங்களை வகுத்து செம்மையாகத் தயாரித்து கணிப்பொறியில் எத்தனை முறை வேண்டுமானாலும் பயன்படுத்தலாம். ஆயிரக்கணக்கானவர்க்கு இத்தகைய பாடங்கள் பயன்படுத்துவதற்கு அமையும். இத்தகைய பாடங்களை ஒரு துறை நிபுணத்துவம் பெற்ற ஆசிரியர்கள் உருவாக்குகின்றனர். அவைகளில் திறமையானது ஆயிரக் கணக்கானோருக்கு உலகம் முழுவதும் பயன்படும்.

11.7.3. கணிப்பொறி வழி கற்றல்-கற்பிப்பித்தல் கருவி உருவாக்குவதற்கான கணிப்பொறி மொழிகள்

கணிப்பொறி வழி கற்பித்தல் மற்றும் கற்றல் பயன்பாடுகளுக்கான மென்பொருள்களை உருவாக்குவதற்கு என்றே கணிப்பொறி அமைப்பு மொழிகள் உள்ளன. IBM – (Internet Bussiness Machine) என்ற நிறுவனம் ஊடிரசளந றுசவைநச என்ற வழியமைப்பு மொழியை இதுபோல் Control Data Corporation என்ற நிறுவனமும் TW என்ற மொழியை ளுவயசமறநயடவாநச என்ற நிறுவனமும் Course Writer என்ற மொழியை Starkwealther என்ற நிறுவனமும் தயாரித்து வெளியிட்டுள்ளன.

கணிப்பொறி வழி கற்றல்-கற்பித்தலுக்கான பாடங்களை உருவாக்க மேற்கூரிய கணிப்பொறி மொழிகள் உதவுகின்றன. இவைகள் ஏனைய கணிப்பொறி மொழிகளை விட எளிமையானவை. இருப்பினும் ஒரு கணிப்பொறி மொழி கற்பதற்கான அதை பயன்படுத்துபவர்களுக்கான திறன் பெற்றவர்க்ள மட்டுமே இவைகளை பயன்படுத்த இயலும். எனவே இவைகளைக் கொண்டு பாடங்களை உருவாக்குபவர்கள் இக்கணிப்பொறி மொழியில் தோ;ச்சி பெறுவது அவசியம் ஆகும். மேலும் வழியமைப்பு திறன் கொண்டவர்களாகவும் இருத்தல் அவசியம்.

11.7.4. கணிப்பொறி வழி கற்றல்-கற்பித்தலுக்கான கருவிகள் உருவாக்குவதற்கான பயனீட்டு மென்பொருள்கள் (packeges)

பொதுவாக வழியமைப்பு மொழிகளை பயன்படுத்தி ஒரு திட்டங்களை உருவாக்குகின்றபொழுது திட்டங்களை செயல்படுத்துவதற்கான முறைமைகளில் தோ;ச்சி பெற்றிருக்க வேண்டும். உதாரணமாக பத்து சொற்களை கணிப்பொறிக்கு தந்து அகரவரிசை பயன்படுத்துவதற்காக கொள்வோம். இப்பணியை செய்ய வழியமைப்பு மொழியில் தோ;ச்சி மட்டும் போதாது. அகரவரிசைப் படுத்துவதற்கான முறைகளை தொந்தவராகவும் இருத்தல் வேண்டும். அவ்வாறு இருந்தால் தான் அந்த முறைகளையும், செயல்பாடுகளையும் கணிப்பொறிக்கு கட்டளைகளாக செயல் படுத்த இயலும். ஆனால் பயனீட்டு மென்பொருள்களில் இதுபோன்ற செயல்பாடுகளுக்கு தேவையானவை தயார் நிலையில் உள்ளன. அவைகளைப் பயன்படுத்தி கொள்வதற்கான சாதாரண திறமைகள் இருந்தால் போதுமானது. எனவே, கணிப்பொறி வழி கற்றல் கற்பிப்பதற்கான மென்பொருள்கள் உருவாக்க கணிப்பொறி மொழிகளை பயன்படுத்த முனைவோர் பாடத் திட்டங்களை மட்டுமல்லாது கணிப்பொறி மொழியிலும் தேர்ச்சிப் பெற்றவராக இருத்தல் வேண்டும். இது ஒரு மிகப்பொpய தடையாகும். இத்தடைகளை அகற்றவும், இப்பணிகளை எளிமைப் படுத்தவும் பல பயனீட்டு மென்பொருள் (Utility Soft Ware) வெளிவருகின்றன. இவைகளில் பாடங்களை கணிப்பொறி வழி

தயார்பபதற்கான அனைத்து வசதிகளும் தயார் நிலையில் உள்ளன. இத்தகைய பயனீட்டு மென்பொருள்களை எப்படி பயன்படுத்திக் கொள்வது என்ற பயிற்சி மட்டுமே போதுமானது. இவைகளை எளிதாகக் கற்றுக் கொள்ளலாம். கணிப்பொறி பயிற்சி இல்லாதவாகள் எளிமையாகவும், விரைவாகவும் கற்றுக் கொள்ள வசதி உள்ளது. Story Board மற்றும் Copy Write போன்றன கணிப்பொறி வழி கற்றல்-கற்பித்தலுக்கான கருவிகள் உருவாக்குவதற்கு பயன்படுத்தப்படும் பயனீட்டு மென்பொருள்கள் ஆகும். இவைகள் தவிர Light Source என்ற மென்பொருள்களும் உள்ளது.

11.7.5. கணிப்பொறி வழி கற்றல்-கற்பித்தலுக்கான கருவியில் நகரும் படங்கள் (animation)

ஊஞஐ மென்பொருள் உருவாக்கத்தில் நகரும் படங்கள் அல்லது சலனப்படங்கள் மற்றும் வீடியோ காட்சிகள் போன்றவைகளை அமைப்பது மென்பொருளின் தரத்தை மேம்படுத்தும். இதற்காக சில மென்பொருள்கள் உள்ளன. அவற்றுள் ஏஐனுநுடீ றுடீசுமுளூ என்பது ஒன்றாகும். இம்மென்பொருளில் சலனப்படங்களை உருவாக்கவும் கையாளவும் தேவையான வசதிகள் உள்ளன.

2.17.6. கணிப்பொறி வழி கற்றல்-கற்பித்தலுக்கான கருவி உருவாக்கத்தில் மனித நுண்ணறிவின் பங்கு

கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. மனிதனைப் போன்று சிந்தித்து செயல்படக் கூடிய கணிப்பொறிகளை உருவாக்க அறிஞர்கள் முனைற்துள்ளனர். கணிப்பொறியை மனிதரைப் போல் உருவாக்கவும், சூழலுக்கு ஏற்ப பணியாற்றவும் அனுபவத்தில் புதிய செய்திகளை சுயமாக கற்கவும் தேவையான திறமைகளை உருவாக்கி செயல்படுமாறு உருவாக்குகின்றன. இத்துறைக்கு செயற்கை நுண்ணறிவு (artificial intelligence) என்று பெயர். இதை ஞஐ என்று சுருக்கமாக அழைப்பர். இத்துறையின் ஒரு அங்கமாக இயற்கை மொழியாய்வு (Natural Language Processing (NLP) என்ற துறை உள்ளது. ஒரு மனிதன் மற்றவர்களிடம் உரையாடுவது போல் மனித மொழியில் கணிப்பொறியுடன் உரையாட தேவையான வசதிகளை இத்துறை செய்கிறது. இந்த ஆய்வு வெற்றி பெறுமேயானால் கணிப்பொறிக்கு கட்டளையிட கணிப்பொறி மொழி என்ற ஒன்று தேவையில்லை. மனித மொழியில் கணிப்பொறிக்கு கட்டளை தரலாம். அதை கணிப்பொறி புரிந்துக் கொண்டு செயலாற்றும். அதுபோல் கணிப்பொறியின் விடையை மனித மொழியில் விளக்கமாகத் தரும். இத்தகைய அமைப்புகள் மனிதனுக்கும், கணிப்பொறிக்கும் இருக்கின்ற உறவை எளிமைப்படுத்தும். மேலும் துறை ஒருங்கு (Expert System)

(ES)) என்பது கணிப்பொறி எழுதியுள்ள மிகப்பெரிய பிரிவாகும். கணிப்பொறியை ஒரு துறையின் வல்லுநரைப் போன்று பயன்படுத்த இவ்வாய்வுகள் உதவும்.

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்கான மென்பொருள்களை உருவாக்கும் பொழுது மேற்கூறிய மொழி நுட்பங்கள் மனித நுண்ணறிவு இயற்கை மொழி ஆய்வு துறை ஒருங்கு போன்றன சேர்க்கப்படுகின்றன. செயற்கை நுண்ணறிவுடன் கூடிய கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்கான மென்பொருளை ACAI என்று அழைக்கிறார்கள். இதை ITS (Intelligence Tutorial System) என்றும் அழைப்பர்.

11.7.7.கணிப்பொறி வழி மொழி கற்றல், கற்பித்தலின் எல்லைகள்

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் கணிப்பொறி சிறந்து பங்காற்றுகிறது. ஆனால் மொழி கற்றல் கற்பித்தல் போன்ற செயல்களில் எண்ணற்ற நுணுக்கங்கள் தேவைப்படுகின்றன. இதில் கணிப்பொறி ஓரளவுதான் பங்காற்ற முடியுமே தவிர பயிற்று விக்கும் ஆசிரியருக்கு நிகராக அமையும் அல்லது அமையாது. மொழி கற்பிப்பதற்கு கணிப்பொறி துணை கருவியாக மட்டும் பயன்படும். இருப்பினும் இத்தகைய மென்பொருள் உருவாக்குகின்ற பொழுது எது இயலுமோ அதில் உறுதியும், திடமான சிந்தனையும் வேண்டும். மொழி கற்றல் கற்பித்தலில் கணிப்பொறியில் செய்யத் தரும் பயிற்சிகள் கீழேப் பட்டியலிடப் பட்டுள்ளன:

- கணிப்பொறி வழி உரையாடல் பயிற்சி அளித்தல்.
- மொழியில் கோடியிட்ட இடங்களை நிரப்புதல் போன்ற வினாத்தாள்களின் பயிற்சி.
- ஒரு சொல்லைத் தந்து அதற்கான உச்சரிப்பு.
- வாக்கிய வகைளை மாற்றி அமைத்தல்.
- இரு சிறு வாக்கியங்களை இணைத்து ஒரு வாக்கியமாக உருவாக்குதல்.
- சொல்லின் ஒருமை, பன்மை, எதிர்பதங்கள் போன்றவற்றை கற்று கொள்வதற்கான பயிற்சிகள்.
- சொற்களை பிரித்து எழுதுதல்
- சொற்களை சோ;த்து எழுதுதல்
- ஒரு சொல்லின் இலக்கணக் கூறினைக் கொடுத்து மற்ற இலக்கணக் கூறுகளைக் கேட்டல்.

மேற்கூறிய பட்டியலில் கணிப்பொறி வழி கற்றல் கற்பித்தலுக்கான சில வாய்ப்புகள் கொடுக்கப்பட்டுள்ளன. இதுபோன்று எளிமையான பயிற்சிகளை தருவதற்கு மென்பொருள்களை உருவாக்கி பயன்படுத்தலாம்.

11.7.8.கணிப்பொறி வழி தமிழ் கற்றல்-கற்பித்தலுக்கான மென்பொருள்கள்

கணிப்பொறி வழி கற்pபத்தலுக்கு எண்ணற்ற மென்பொருள்கள் உள்ளன. அறிவியல் தொழில் நுட்பம் போன்ற பாடங்களை பயிற்றுவிப்பதற்கு உலகளவில் பல்லாயிரக்க கணக்கான மென்பொருள்கள் உருவாக்கப்பட்டு பயன்படுத்தப்பட்டு வருகின்றன. அதுபோல் மொழி கற்றல் கற்பித்தலுக்குப் பல மென்பொருள்கள் உள்ளன. கணிப்பொறி வழி தமிழ் கற்றல் கற்பித்தலுக்கு மென்பொருள் உருவாக்கம் தொடங்கப்பட்டு நிகழ்ந்துவருகின்றது.

சென்னையைச் சார்ந்த Learn Fun என்ற நிறுவனம் தமிழுக்காக மென்பொருள் ஒன்றை உருவாக்கி உள்ளது. இம்மென்பொருளில் தமிழ் எழுத்துக்களை எழுத, பழக வசதிகள் உள்ளன. இம்மென்பொருள் மொழி விளையாட்டு மென்பொருள் வகையைச் சேர்ந்தது. தமிழை ஆரம்ப நிலையில் கற்பவர்கள் வல்லினம், மெல்லினம், குறில், நெடில் போன்ற அடிப்படைக் கூறுகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள் மிகவும் உதவுகிறது. சென்னையைச் சார்ந்த Soft View என்ற நிறுவனம் அமுதம் என்ற மெனபொருளை உருவாக்கியுள்ளது. இம்மென்பொருள், திருக்குறள், ஆத்திச் சூடி போன்றவைகளை எளிமையாகக் கற்று தருகிறது. சிங்கப்பூர் தேசிய கல்விக் கழகம் தமிழில் பயிற்றுவிப்பதற்காக மென்பொருள் ஒன்றை உருவாக்கியுள்ளது. ஒருமை, பன்மை, எதிர்ப்பதம் போன்றவைகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள் உதவுகிறது. சிங்கப்பூரைச் சார்ந்த தமிழாசிரியர் கழகம் தமிழைப் பயிற்றுவிப்பதற்காக சில மென்பொருளை வெளியிட்டுள்ளன. அவைகளைக் கொண்டு தமிழில் உள்ள இரண்டு எழுத்துச் சொற்கள், மூன்று எழுத்துச் சொற்கள் போன்றவைகளை எளிமையாக கற்கலாம். மேலும்; வெளிப்படும் படத்திலிருந்து அப்படத்தில் உள்ள பொருளை அடையாளம் காண்டு பயிற்சிப் பெற சில வசதிகள் இம்மென்பொருளில் உள்ளன. மேலும் ஒன்றியிலிருந்து ஆயிரம் வரை எண்கள் எழுதவும், படிக்கவும் சொல்லித் தருகிறது. மேற்கண்ட மென்பொருள் தயாரிக்கும் பணி அமரர் பேராசிரியார் கோவிந்தசாமி தலைமையில் சிங்கப்பூரில் நடைப்பெற்றது. அமெரிக்காவைச் சார்நத லிண்ட்ஹோம் என்பவர் தமிழ்ச் சொற்கோவைகளை பயிற்றுவிப்பதற்காக மென்பொருள் ஒன்றை வெளியிட்டுள்ளார். பென்சில்லேனியா பல்கலைக் கழகத்தில் பேச்சுத் தமிழைப் பயிற்றுவிப்பதற்காக மென்பொருள் உருவாக்க முனைந்துள்ளனர். அமெரிக்காவில் உள்ள சிக்காக்கோ நகரிலுள்ள நிறுவனம் ஒன்று Visual Tamil என்ற மென்பொருளை உருவாக்கியுள்ளது. இம்மென்பொருள் தமிழ் கலை, பண்பாடு ஆகியவற்றைக் கற்பிக்கிறது. இது போன்று ஆங்காங்கே கணிப்பொறி வழி தமிழில் கற்பித்தல் கற்றலுக்கு பல மென்பொருள்கள் உருவாக்கப்படுகின்றன.

11.7.9. கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் இணையத்தின் பங்கு

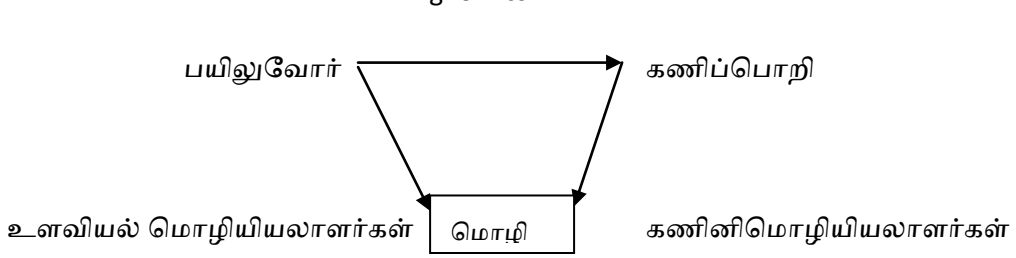
ஒன்றுக்கு மேற்பட்ட கணிப்பொறிகள் ஒன்றோடு ஒன்று இணைக்கப்படுகின்றன. இத்தகைய அமைப்பிற்கு கணிப்பொறி வலைப்பின்னல் (Computer Net work) என்று பெயர். இதனால் ஒரு கணிப்பொறியில் பதிவு செய்யப் பட்டுள்ள மென்பொருள் மற்றும் பதிவுகளை கணிப்பொறிகளில் பெறலாம். இது போன்று உலகளாவிய நிலையில் உள்ள கணிப்பொறிகளை ஒன்றிணைத்து தகவல் பரிமாற்றம் செய்து கொள்ள உதவும் அமைப்பிற்கு இணையம் என்று பெயர். இதன் பயனாய் இணையத்தில் உள்ள ஒரு கணிப்பொறியில் பதிவு செய்யப்படும் செய்திகளை உலகில் உள்ள எந்த கணிப்பொறியிலிருந்தும் பெறலாம். கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்கான மென்பொருள் இணையம் ஏற்றம் செய்யப்பட்டுள்ளன. இம்மென்பொருள்களை செயல்படுத்தி மொழிகளை கற்பிக்கலாம். பென்சில்லேனியா பல்கலைக் கழகத்தில் WALTT – Web Aided Language Teaching and Learning அல்லது Training என்ற திட்டம் செயல்படுத்தப் பட்டுள்ளது.

இணையத்தின் வழி பேச்சு தமிழை பயிற்றுவிப்பதற்கு இது உதவுகிறது. மேலும் தமிழ் மொழியை கற்பிப்பதற்கு சிங்கப்பூரில் உருவாக்கப்பட்டுள்ள பல இணையத் தளங்கள் செம்மையாக உள்ளன. தமிழ்ப்பல்கலைக் கழகத்தின் அங்கமாக தமிழ் இணையப் பல்கலைக்கழகம் செயல்படுகிறது. இணையப் பல்கலைக்கழகம் தமிழ் பாடத்தில் சான்றிதழ் வகுப்புகளையும், பட்டய வகுப்புகளையும், தொலைத்தூர கல்வி திட்டத்தின் கீழ் இணையம் வழி நடத்துகிறது. அயல் நாட்டில் உள்ள தமிழர்கள் மற்றும் இந்திய மாநிலங்களில் உள்ள தமிழர்கள் தமிழ் மொழியை இதன் வழி கற்கலாம். இப்பல்கலைக் கழகம் தமிழ் மொழி கலாச்சாரம், பண்பாடு போன்றவற்றை இணையம் வழி கற்பிக்கிறது. மேலும் தமிழ் மொழியை கற்றுக் கொள்வதற்கு உறுதுணையாக அமையும் அகராதிகள், இலக்கியங்கள் போன்றவைகளை இந்த இணையத்தின் மூலம் இலவசமாக கற்கலாம். இது போன்று தமிழ்மொழி கற்றலுக்கும், கற்பித்தலுக்கும் இணையத்தில் பல வாய்ப்புகள் உள்ளன.

11.7.10. மொழி கற்றல் மென்பொருள் உருவாக்கத்தில் பிற துறையினரின் பங்கு

இத்தகைய மென்பொருளை உருவாக்க மொழி வல்லுநர் உளவியல் மொழியியாளர்கள் (Psycho-linguistics experts) கணினி மொழியியல் வல்லுநர்கள் (Computational Linguistics experts) ஆகியோரின் ஒத்துழைப்பு அவசியமாகும். மொழியியல் வல்லுநர்கள் கணிப்பொறிக்கும், மொழிக்கும் உள்ள உறவை ஆராய்கின்றனர். உளவியல் மொழியாளர்கள் பயில்வோருக்கும், மொழிக்கும் உள்ள உறவை ஆராய்கின்றனர். மேலும் கணிப்பொறிக்கும், பயில்வோருக்கும்

உள்ள உறவை ERGNOMICS என்ற துறையினர் ஆராய்கின்றனர். இத்தொடர்பு கீழே கண்ட கட்டப்படம் விளக்குகிறது.



கணிப்பொறி வழி கற்றல், கற்பித்தலில் மேற்கூறியவாறு கணிப்பொறி பல வகைகளில் பயன்படுகிறது.

11.8. சுருக்க உரை

முன்னர் கூறப்படாத சில மொழிக்கருவிகள் இங்குச் சுருக்கமாக விளக்கப்பட்டுள்ளன. ஒலிபெயர்ப்பி உருவாக்குதல், எழுத்துப் பெயர்ப்பி உருவாக்குதல், எழுத்துணரி உருவாக்குதல், சொற்களை அகரவரிசைப் படுத்துதல், சொல்லாய்வி அல்லது சொல்லாளர் உருவாக்குதல், மொழிக்கற்றல் கற்பித்தல் உருவாக்குதல் என்பன இதில் அடங்கும். எல்லா வசதிகளும் கொண்ட சொல்லாய்வி உருவாக்குவதும் கற்றல்-கற்பித்தல் கருவிகள் உருவாக்குவதும் எழுத்துணரி உருவாக்குவதும் கடின மான செயல்கள். சொல்லாய்வி தமிழுக்குப் பலரால் உருவாக்கப் பட்டுள்ளது. தெய்வசுந்தரம் அவர்களின் சொல்லாளர் மென் தமிழ் சிறப்பானதாகும். கற்றல்-கற்பித்தல் கருவிகள் பலரால் உருவாக்கப் பட்டுள்ளது. வெளிநாட்டுத் தமிழர்களால் குறிப்பாகச் சிங்கப்பூர் தமிழர்களால் சிறந்த கற்றல்-கற்பித்தல் கருவிகள் உருவாக்கப்பட்டுள்ளன. மொழியியலார் இராசாராம் அவர்களின் கற்றல்-கற்பித்தல் கருவி குறிப்பிடத் தகுந்ததாகும். கிருஷ்ணமூர்த்தி அவர்கள் தமிழ் எழுத்துணரி உருவாக்குவதில் தலைசிறந்தவர் ஆவார். கூகிள் தமிழுகு ஓர் எழுத்துணரியை உருவாக்கித் தந்துள்ளது.

Language in India www.languageinindia.com Vol. 19 Issue 11, Nov 2019

இயல் 12

தமிழின் தொழில்நுட்ப வளர்ச்சி

12.1. அறிமுகம்

கணினி மொழியியல் என்பது ஒரு கணினிசார் கண்ணோட்டத்தில் இயற்கையான மொழியின் புள்ளிவிவர அல்லது விதி அடிப்படையிலான மாதிரி, அத்துடன் மொழியியல் கேள்விகளுக்குப் பொருத்தமான கணினி அணுகுமுறைகளைப் பற்றிய ஆய்வு ஆகியவற்றுடன் தொடர்புடைய ஒரு இடைநிலை துறையாகும். பாரம்பரியமாக, கணினி மொழியியல் இயற்கை மொழியின் ஆய்விற்குக் கணினிகளைப் பயன்படுத்துவதில் நிபுணத்துவம் பெற்ற கணினி அறிவியலாளர்களால் நிகழ்த்தப்பட்டது. இன்று, கணினி மொழியியலாளர்கள் பெரும்பாலும் இடைநிலைக் குழுக்களின் உறுப்பினர்களாகப் பணியாற்றுகிறார்கள்; இதில் வழக்கமாக மொழியியலாளர்கள், இலக்கு மொழியில் வல்லுநர்கள் மற்றும் கணினி அறிவியலார்கள் உள்ளனர். பொதுவாக மொழியியலாளர்கள், கணினி அறிவியலார்கள், செயற்கை நுண்ணறிவு வல்லுநர்கள், கணிதவியலாளர்கள், தர்க்கவியலாளர்கள், தத்துவவாதிகள், புலனறிவு அறிவியலாளர்கள், புலனறிவு உளவியலாளர்கள், உளவியலாளர்கள், மானுடவியலாளர்கள் மற்றும் நரம்பியல் அறிவியலார்கள் ஆகியோரின் ஈடுபாட்டை கணினி மொழியியல் ஈர்க்கிறது. கணினி மொழியியல் கோட்பாட்டு மற்றும் பயன்பாட்டுக் கூறுகளைக் கொண்டுள்ளது. கோட்பாடுக் கணினி மொழியியல் கோட்பாட்டு மொழியியல் மற்றும் புலனறி அறிவியலில் உள்ள சிக்கல்களை மையமாகக் கொண்டுள்ளது, மேலும் பயன்பாட்டுக் கணினி மொழியியல் மனித மொழிப் பயன்பாட்டை மாதிரியாக்குவதன் நடைமுறை விளைவுகளை மையமாகக் கொண்டுள்ளது. கணினி மொழியியல் சங்கம் (Association for Computational Linguistics) கணினி மொழியியலை இவ்வாறு வரையறுக்கிறது: ஒரு கணக்கீட்டு கண்ணோட்டத்தில் கணினி மொழியியல் மொழியின் அறிவியல் ஆய்வு ஆகும். கணினி மொழியியலாளர்கள் பல்வேறு வகையான மொழியியல் நிகழ்வுகளின் கணினிசார் மாதிரிகளை வழங்க ஆர்வமாக உள்ளனர்.

இயற்கை மொழி ஆய்வு (Natural Language Processing (NLP/என்.எல்.பி.) மற்றும் கணிமொழியியல் (computational Linguistics) ஆய்வின் வளர்ச்சியைக் குறிப்பிடும் போது நாம் நேரிடுவது இவ்வளர்ச்சியின் வெளிப்பாடாய் மொழி ஆய்வில் ஏற்பட்ட மாற்றங்களும் வளர்ச்சிகளும் ஆகும்; குறிப்பாக இயற்கைமொழி ஆய்வு மற்றும் கணினி மொழியியலின் காரணமாக தமிழ் மொழி பெற்ற தொழில் நுட்பவளர்ச்சியை தான் நாம் இங்கு கருத்தில் கொள்ளவேண்டியது.

இயற்கை மொழி ஆய்வுக்கும் கணினி மொழியியலுக்கும் சிலர் வேற்றுமை பாராட்டுவர். சிலர் இரண்டையும் ஒன்றாகக் கருதுவர். இரண்டிற்குமான வேறுபாட்டைக் கணக்கிடுவது நமது நோக்கமல்ல. இரண்டும் தமிழுக்குச் செய்த பங்களிப்புதான் முக்கியம்.

"Technological Development of Tamil" (Rajendarn 2017) என்ற கட்டுரை இயற்கை மொழி ஆய்வு மற்றும் கணிமொழியியல் பின்னணியி தமிழ் அடைந்த தொழில் நுட்ப வளர்ச்சி குறித்து விளக்குகின்றது. இதன் அடிப்படையில் இயற்கை மொழி ஆய்வு மற்றும் கணிமொழியியல் ஆய்வின் வளர்ச்சியின் வெளிப்பாடாக தமிழ் பெற்ற தொழில் நுட்ப வளர்ச்சிகளையும் சவால்களையும் இக்கட்டுரை விளக்குகின்றது. இக்கட்டுரை எழுத 287 நோக்க்கீடுகள் பயன்படுத்தப்பட்டுள்ளன. இது முழுமையல்ல. மேலும் மேப்படுத்தப்படவேண்டும்.

12.2. சவால்கள்களும் இடையூகளும்

தமிழ் மொழியின் தொழி நுட்ப வளர்ச்சி டிஜிட்டல்/இலக்க மயமாக்கல், யூனிகோட் எழுத்துருக்களுக்கு மாறுதல், ஏற்கனவே கிடைத்துள்ள டிஜிட்டல் மயமாக்கப்பட்ட பனுவகளை அல்லது உரைகளின் எழுத்துரு மாற்றி மூலம் யூனிகோட் வடிவத்தில் மாற்றுதல், தமிழ் கணிப்பீட்டிற்கும் தமிழ்த் தொழில்நுட்ப மேம்பாட்டுக்கான பிற முயற்சிகளுக்கும் வேண்டி மொழியியல் அல்லது மொழித் தொழில்நுட்ப வளங்களை உருவாக்குதல் என்பன தமிழ் நேரிடுகின்ற சவால்களும் இடையூறுகளில் சிலவாகும். இடையூறு அல்லது இழப்பு மட்டத்தில் தமிழும் ஆங்கிலம் போன்ற மொழிகலந்த கலப்பு குறியீட்டு முறை (mixed coding), துய்மையற்ற மற்றும் இலக்கணப் பிழையான மொழி வெளிப்பாடுகள், கொச்சை மொழிப் பயன்பாடுகள் போன்ற பல்வேறு வகையான சிதைவுகளை டிஜிட்டல்/இலக்க ஊடகங்களில் தமிழ் எதிர்கொள்கிறது. இந்த நடைமுறை தமிழ் மொழியில் பண்பாடு மற்றும் பல்வேறு தனித்துவமான அம்சங்களை இழக்க வழிவகுக்கும்.

டிஜிட்டல்மயமாக்கல் என்பது தகவல்களை டிஜிட்டல் வடிவமாக மாற்றும் செயல்முறையாகும். இந்த வடிவமைப்பில், தகவல் தனியாக முகவரித்தர இயலும் (பெரும்பாலும் பைட்டுகள் என்று அழைக்கப்படுகின்ற பன்முக பிட்களில் இருக்கும்) தரவின் (பிட்கள் என அழைக்கப்படுகிற) தனித்தனி அலகுகளாக ஒழுங்கமைக்கப்பட்டுள்ளது. கணினிகள் மற்றும் கணினி திறன் கொண்ட பல சாதனங்கள் (டிஜிட்டல் கேமராக்கள் மற்றும் டிஜிட்டல் கேட்டல் கருவி போன்றவை) செயலாக்கக்கூடிய இருமத் தரவு இது. டிஜிட்டல் இந்தியா முயற்சி சில அற்புதமான திட்டங்களை அறிமுகப்படுத்தியுள்ளது.

உரை மற்றும் படங்களை இதேபோல் டிஜிட்டல் மயமாக்கலாம்: ஒரு ஒளிவருடி (ஸ்கேனர்) ஒரு படத்தைப் பிடிக்கிறது (இது உரையின் படமாக இருக்கலாம்) மற்றும் அதை பிட்மேப் போன்ற படக் கோப்பாக மாற்றுகிறது. ஒளி எழுத்துணரி (Optical Chracter Recognizer (OCR)) நிரல் ஒவ்வொரு ஒரு உரை படத்தையும் நெடுங்கணக்கு எழுத்து அல்லது எண் இலக்கத்தை அடையாளம் காணும் பொருட்டு ஒளியுள்ள மற்றும் இருண்ட பரப்புகளுக்கான பகுப்பாய்வு செய்கிறது, மேலும் ஒவ்வொரு எழுத்தையும் ASCII குறியீடாக மாற்றுகிறது.

கேட்டல் மற்றும் காண்டல் (ஆடியோ மற்றும் வீடியோ) டிஜிட்டல் மயமாக்கல் அனலாகிலிருந்து டிஜிட்டலாக மாற்றும் செயல்முறைகளில் ஒன்றைப் பயன்படுத்துகிறது, இதில் தொடர்ச்சியாக மாறக்கூடிய (அனலாக்) சமிக்ஞை, அதன் தேவையான உள்ளடக்கத்தை மாற்றாமல், பல நிலை (டிஜிட்டல்) சமிக்ஞையாக மாற்றப்படுகிறது. மாதிரி செயல்முறை ஒரு அனலாக் அலைவடிவத்தின் வீச்சை (சமிக்ஞை வலிமை) சம இடைவெளியில் உள்ள நேரக் குறிப்பான்களில் அளவிடும் மற்றும் மாதிரிகளை டிஜிட்டல் தரவுகளாக உள்ளீடு செய்வதற்கான எண் மதிப்புகளாக உருப்படுத்தம் செய்யும்.

தகவலை டிஜிட்டல் செய்வது பாதுகாப்பதையும் அணுகுவதையும் பகிர்வதையும் எளிதாக்குகிறது. எடுத்துக்காட்டாக, ஒரு அசல் வரலாற்று ஆவணம் அதன் இருப்பிடத்தைப் பார்வையிடும் நபர்களுக்கு மட்டுமே அணுகக்கூடியதாக இருக்கலாம், ஆனால் ஆவண உள்ளடக்கம் டிஜிட்டல் மயமாக்கப்பட்டால், அதை உலகளவில் மக்களுக்கு கிடைக்கச் செய்யலாம். வரலாற்று ரீதியாகவும் பண்பாட்டு ரீதியாகவும் குறிப்பிடத்தக்க தரவுகளை டிஜிட்டல் மயமாக்குவதற்கான போக்கு அதிகரித்து வருகிறது. தமிழ் இணையக் கல்விக்கழகம், மதுரை திட்டம் (project Maduai) மற்றும் பிற நிறுவனங்கள் தமிழ் இலக்கியப் படைப்புகளை (சங்கத் தமிழ் இலக்கிய நூல்கள், இடைக்கால மற்றும் நவீனகால தமிழ் இலக்கியப் படைப்புகள் மற்றும் பிற முக்கிய படைப்புகளை) டிஜிட்டல் மயமாக்கியுள்ளன.

தமிழ் உள்ளீட்டு முறைகள் தட்டச்சுப்பொறி அல்லது கணினி விசைப்பலகையைப் பயன்படுத்தி தமிழ் மொழி எழுத்துக்களை உள்ளீடு செய்ய உருவாக்கப்பட்ட வெவ்வேறு ஒழுங்கமைப்புகளைக் குறிக்கின்றன. அழகி மற்றும் என்.எச்.எம் (New Horizon Meida NHM)) எழுத்தாளர் போன்ற பல நிரல்கள் தட்டச்சு செய்வதற்கான நிலையான மற்றும் ஒலிப்பு வகை தளவமைப்புகளை வழங்குகின்றன.

தமிழ் மொழிக்கான தற்போதைய யூனிகோடில் பல சிக்கல்கள் உள்ளன. தமிழிற்கான தற்போதைய யூனிகோட் தரநிலை பின்வரும் காரணங்களால் கணினிகளில் தமிழின் திறமையான மற்றும் பயனுள்ள பயன்பாட்டிற்கு போதுமானதாக இல்லை என்று கருதப்படுகிறது:

(1) 247 தமிழ் எழுத்துக்களில் 31 மட்டுமே யூனிகோட் குறியீடு நிலைகளைக் கொண்டுள்ளன. இந்த 31 எழுத்துக்களில் 12 உயிரெழுத்துக்கள், 18 உயிர்-மெய்கள் மற்றும் ஒரு ஆய்தம் ஆகியவை அடங்கும். ஐந்து கிரந்த உயிரெழுத்து-மெய் சேர்க்கைகளுக்கு யூனிகோட் தமிழில் குறியீடு இடம் வழங்கப்படவில்லை. பிற தமிழ் எழுத்துக்கள் தனி மென்பொருளைப் பயன்படுத்தி பெறப்பட வேண்டும். தற்போதைய யூனிகோட் தமிழில், 10% தமிழ் எழுத்துக்களுக்கு மட்டுமே குறியீடு இடம் வழங்கப்பட்டுள்ளது. பொதுவான உரை பரிமாற்றத்தில் பயன்படுத்தப்படும் 90% தமிழ் எழுத்துக்களுக்கு குறியீடு இடம் வழங்கப்படவில்லை.

(2) தற்போதைய யூனிகோட் தமிழில் விடப்பட்டிருக்கும் உயிரெழுத்து-மெய் எழுத்துகள் ஏ, பி, சி, டி போன்ற ஆங்கிலத்திற்கான எழுத்துக்கள் போன்று எளிய எழுத்துக்கள் ஆகும். உயிரெழுத்து-மெய் எழுத்துகள் கிளிஃப்கள், அல்லது பிணைப்புகள் (ligatures) அல்லது யூனிகோடில் கருதப்படும் இணைந்த எழுத்துக்கள் அல்ல. க, கா, கி, கீ போன்றவை தமிழின் எழுத்துக்கள்.

(3) எந்தவொரு எளிய தமிழ் உரையிலும், உயிர்-மெய் 64 முதல் 70% வரை உருவாக்கப்படுகிறது; உயிரெழுத்துகள் 5 முதல் 6% வரையிலும், மெய் 25 முதல் 30% வரையிலும் உருவாக்கப்படுன்றன. உயிர்-மெய் போன்ற உயர் நிகழ்வெண் எழுத்துக்களை கிளிஃப்களாக உடைப்பது மிகவும் திறமையற்றது.

(4) இந்த வகை குறியமாக்கத்திற்கு ஒரு எழுத்தை உணர ஒழுங்கமைப்பு (rendering) இயந்திரம் தேவைப்படுகிறது. இந்த வகைக் கணினியாக்கம் (computing) தமிழில் கணினி மென்பொருள் மேம்பாடுகள், தேடல் மற்றும் வரிசையாக்கம் மற்றும் தமிழில் இயற்கை மொழி ஆய்வு போன்ற பயன்பாடுகளுக்கு ஏற்றதல்ல. இது கூடுதல் நேரத்தையும் இடத்தையும் பயன்படுத்துகிறது. இது கணினிச் செயல்முறையை மிகவும் திறமையற்றதாக ஆக்குகிறது. அத்தகைய பயன்பாடுகளுக்கு ஆங்கிலம் போன்று ஒரு மொழியின் அனைத்து எழுத்துக்களும் குறியமாக்கத்தில் குறியீட்டு நிலைகளைக் கொண்டுள்ள நிலை-1 (Level-1) செயல்படுத்தம் தேவை.

(5) இந்தக் குறியமாக்கம் இஸ்கீ/ஐ.எஸ்.சி.ஐ.ஐ. (Indian Script Code for Information Interchange (ISCII) (1988) அடிப்படையாகக் கொண்டது. எழுத்துக்கள் இயற்கையான வரிசையில் இல்லை. இயற்கையான வரிசையில் அவற்றை ஒழுங்கமைக்க ஒரு சிக்கலான சேர்ந்துவருகை வழிமுறை (collocational algorithm) தேவைப்படுகிறது.

(6) பல குறியீட்டுப் புள்ளிகளை பயன்படுத்தி தனி எழுத்துகள் (characters) எழுத்துக்கள் பெறப்படுகின்றன. பல்குறியீடுப் புள்ளிகள் பாதுகாப்புப் பாதிப்புகள், பொருள்மயக்கமான சேர்க்கைகளுக்கு வழிவகுக்கும் மற்றும் இயல்பாக்கலின் பயன்பாடு தேவைப்படும்.

12.3. வளங்கள் மற்றும் கருவிகள் (Resources and tools)

12.3.1. மொழி வளங்கள்: வளங்கள், தரவு மற்றும் அறிவுத் தளம்

இயற்கைமொழி ஆய்வுச் சூழலில், தரவுத்தொகுதி (corpus) என்பது உரையின் பெரிய தொகுப்பாகும். உரைத் தரவுத்தொகுதி, பேச்சுத் தரவுத்தொகுதி, படத் தரவுத்தொகுதி போன்ற பல்வேறு வகையான தரவுத்தொகுதிகள் உள்ளன.

12.3.1.1. உரை தரவுத்தொகுதிகள் (Text corpora)

தகவல் தொடர்பு மற்றும் தகவல் தொழில்நுட்ப அமைச்சகத்தின் இந்திய மொழிகளுக்கான தொழில்நுட்ப மேம்பாட்டுத் திட்டத்தின் (Technological Development of Indina Lanaguages (TDIL) திட்டத்தின் கீழ், பட்டியலிடப்பட்ட மொழிகளுக்கு 45 மில்லியனுக்கும் அதிகமான சொல் தரவுத்தொகுதையை உருவாக்க மத்திய இந்திய மொழிகள் நிறுவனம் ஒருங்கிணைத்துள்ளது. இந்தத் தரவுத்தொகுதிகள் மாதிரி முறைகளைப் பின்பற்றி உருவாக்கப்பட்டுள்ளன, எனவே இவை சீரான தரவுத்தொகுதிக்கள் ஆகும். அவை தகவல் பரிமாற்றத்திற்காக இந்தியத் தரநிலை குறியம் (Indian Standard code) அல்லது தகவல் பரிமாற்றத்திற்கான இந்திய எழுத்துக் குறியம் (Indian Script Code for Information Interchange (ISCII) வடிவத்தில் கிடைக்கின்றன. இந்த தரவுத்தொகுதிகளை ஒவ்வொரு மொழியிலும் இருபது மில்லியனாக உயர்த்தவும் நிறுவனம் விரும்புகிறது.

லான்காஸ்டர் பல்கலைக்கழகத்துடன் இணைந்து, இந்திய மொழிகளின் மையநிறுவனம் இதை யுனிகோட் வடிவமாக மாற்றியுள்ளது. லான்காஸ்டர் பல்கலைக்கழகத் தரவுத்தொகுதியுடன் இந்த வடிவமைப்பில் உள்ள தரவுத்தொகுதியும் பயனர்களுக்கு கிடைக்கிறது: http://www.emille.lancs.ac.uk/home.htm

தமிழ் இணையக் கல்விக்கழகம் உலகளாவிய தரத்துடன் தமிழ் தரவுதொகுதி வங்கியை உருவாக்குவதை நோக்கமாகக் கொண்டுள்ளது. தற்போது, சங்ககால, இடைக்கால மற்றும் நவீனகாலத் தமிழிலிருந்து 150 மில்லியன் சொற்களின் தரவுத்தொகுதி உருவாக்கியுள்ளது.

இந்திய மொழிகள் தரவுத்தொகுதி முன்முயற்சி கட்டம்- II (Indian Languages Copora Intitiative-Phase II) திட்டத்தின் கீழ், இந்திய அரசின் மின்னணு மற்றும் தகவல் தொழில்நுட்ப அமைச்சகத்தால் (Ministry of Electronics and Information Technology (MeitY) ஊக்குவிக்கப்பட்டு, புது தில்லி ஜவஹர்லால் நேரு பல்கலைக்கழகம் தமிழில் ஒருமொழி தரவுத்தொகுதியைச் சேகரித்துள்ளது. இது திட்டத்தின் இறுதி விளைவாகும்; தோராயமாக பொதுக் களத்தின் 30,000 வாக்கியங்கள் உள்ளன.. மொழிபெயர்க்கப்பட்ட வாக்கியங்கள் இந்திய தரநிலைகளின் பணியகத்தின் (பிஓஎஸ்) (Bureau of Indian Standards (BIS) அடையாள தொகுதியின் படி சொல்வகைப்படுகளுக்காக அடையாளப்படுத்தப் பட்டுள்ளன. இந்தத் தரவுத்தொகுதி பின்வரும் அம்சங்களைக் கொண்டுள்ளது: தனிப்பட்ட அடையாளம், யுடிஎஃப்-8 குறியமாக்கம் மற்றும் உரை கோப்பு வடிவம்.

இந்திய மொழி தொழில்நுட்பப் பெருக்கம் மற்றும் பயன்படுத்தல் மையம் (Indian Language Technology Proliferation & Deployment Centre (ILTPDC/ஐ.எல்.டி.பி.டி.சி) தரவிறக்கம் செய்யக்கூடிய தமிழ்-தெலுங்கு பொது உரை தரவுத்தொகுதியைக் கொண்டுள்ளது. இந்திய மொழியிலிருந்து இந்திய மொழி மொழிபெயர்ப்பு திட்டத்தின் கீழ் தமிழ்-தெலுங்கு இணைத் தரவுத்தொகுதி மனித மொழிபெயர்ப்பாளர்களால் தயாரிக்கப்பட்டுள்ளது. இதில் 3:1 என்ற விகிதத்தில் சுற்றுலா/பயணம் மற்றும் பிறவற்றை உள்ளடக்கிய தரவுத்தொகுதி இணையத்திலிருந்து எடுக்கப்பட்டுளது. மிகவும் எளிமையான, எளிய, சிக்கலான மற்றும் கூட்டு வாக்கிய கட்டமைப்புகளைக் கொண்ட மொத்தம் 1000 வாக்கியங்கள் உள்ளன. வாக்கிய கட்டமைப்புகளில் பெயரெச்சத்தொடர்கள், நிரப்பு எச்சத்தொடர்கள், முற்று எச்சத்தொடர்கள், குறைநிலை எச்சத்தொடர்கள் இணைப்புகள் என்பன அடங்கும். உரை குறியாக்கம் UTF-8 ஆகும் (ILTPDC இணையதளத்தில் தெரிவிக்கப்பட்டுள்ளது).

ஒரு பெரிய அளவு மனிதமுயற்சியால் சொல்வகைபாட்டிற்காக அடையாளப்படுத்தப்பட தரவுத்தொகுதி (515 கே டோக்கன்கள்) சென்னை ஏ.யூ.கே.பி.சி.ஆர்.சி. (Anna University K.B.Chandrsekhar Research Centre (AUKBCRC/எயூகேபிசிஆர்சி) மூலம் தமிழ்க்குக் கிடைக்கிறது. தரவுத்தொகுதி கல்கி கிருஷ்ணமூர்த்தி எழுதிய 20 ஆம் நூற்றாண்டின் புகழ்பெற்ற தமிழ் நாவலான "பொன்னியன் செல்வன்" என்பதை அடிப்படையாகக் கொண்டதாகும். தரவுத்தொகுதி பிஸ் அடையாளத் தொகுதியால் அடையாளப்படுத்தப்பதாகும். பிஸ் அடையாளத் தொகுதி இந்திய

தரநிலைகள் பணியகம் மற்றும் தமிழ் இணையக் கல்விக்கழகம் இவற்றால் அங்கீகரிக்கப்பட்டுள்ள ஒரு படிநிலை அடையாளத்தொகுதியாகும். தரவுத்தொகுதியின் புள்ளிவிவரம்: வாக்கியங்களின் மொத்த எண்ணிக்கை 50,876; சொற்களின் எண்ணிக்கை 5,15,283.

12.3.1.2. பேச்சுத் தரவுத்தொகுதிகள் (Speech corpora)

ஐ.எல்.டி.பி.டி.சி பதிவிறக்கம் செய்ய இயலும் தமிழ் பேச்சு தரவின் (ASR) இருப்பை அறிவிக்கின்றது. இந்த தரவுத்தொகுதி 1000 பேச்சாளர்களின் தமிழ் மொழியின் 62000க்கும் மேற்பட்ட கேட்பு கோப்புகள் உள்ளன. இந்தத் தரவு வேளாண் பொருட்கள் களத்திற்காகத் தயாரிக்கப்பட்டது மற்றும் இந்தத் தரவுத்தொகுதியின் அளவு 5.7 ஜிபி ஆகும்.

இந்தத் தமிழ் பேச்சு அறிதல் தரவுத்தளம் (Speech Recognition Database) தமிழ்நாட்டில் சேகரிக்கப்பட்டது; இது வயது வேறுபாடு (16-20, 21-50, 51+), பாலினம், கிளைமொழிப் பகுதிகள் மற்றும் சுற்றுச்சூழல் (வீடு, அலுவலகம் மற்றும் பொது இடம்) ஆகியவற்றின் படி தேர்ந்தெடுக்கப்பட்ட 450 வெவ்வேறு இயல்புப் பேச்சாளர்களின் குரல்களைக் கொண்டுள்ளது. ஒவ்வொரு பேச்சாளரும் ஓசை இல்லாத சூழலில் ஒரு செய்தி உரையை ஒலிப்பதிவுக் கருவி மூலம் உள்ளடங்கிய ஒலிவாங்கி (மைக்ரோஃபோன்) மூலம் பதிவு செய்தனர். பதிவுகள் ஒலிபெருக்கிப் பதிவில் (ஸ்டூரியோ ரிகார்டிங்க்) உள்ளன மற்றும் பிரித்தெடுக்கப்பட்ட சேனலும் குறிப்பிட்ட கோப்புகளில் சேர்க்கப்பட்டுள்ளன. இதில் ZIP கோப்புகளாக சேமிக்கப்பட்ட கேட்பு (ஆடியோ) கோப்பு, உரை கோப்பு, என்ஐஎஸ்டி கோப்புகள் ஆகியவை அடங்கும். அனைத்து பேச்சுத் தரவும் எழுத்துவடிவமாக்கப்பட்டு வாக்கிய மட்டத்தில் அடையாளப்படுத்தப்பட்டுள்ளன. புள்ளிவிவரம்: வடிவம் மற்றும் செயல்பாட்டு சொற்கள், கட்டளை மற்றும் கட்டுப்பாட்டு சொற்கள், ஒலிப்பு சீரான சொற்றொகுதி, சரியான பெயர்கள், அடிக்கடி 1000 சொற்கள்.

12.3.1.3. இணைத் தரவுத்தொகுதிகள் (Parallel corpora)

“என்டாம்: ஆங்கிலம்-தமிழ் இணை தரவுத்தொகுதி’ (EnTam: An English-Tamil Parallel Corpus) என்ற திட்டத்தின் கீழ் ஒரு தமிழ் தொடர்பான இயற்கைமொழி ஆய்வுக்காகப் பொதுவில் கிடைக்கும் சில வலைத்தளங்களிலிருந்து ஆங்கிலம்-தமிழ் இருமொழியத் தரவு சேகரிக்கப்பட்டுள்ளது. சீரமைக்கப்படாத வலைத் தரவின்மீது தரமான செயலாக்கத் தொகுப்புகள் பிரயோகிக்கப்பட்டு பல்வேறு இயற்கைமொழி ஆய்வுப் பணிகளுக்கு ஏற்றவாறு வாக்கியங்கள் இணையாக வரிசைப்படுத்தப்பட ஆங்கில-தமிழ் இணைத் தரவுத்தொகுதி கிடைக்கின்றது. இந்த இணைத் தரவுத்தொகுதிகள் பைபிள், சினிமா மற்றும் செய்தி களங்களிலிருந்து உரைகளை உள்ளடக்கியது.

ராமமூர்த்தி அறிவித்தபடி சிஐஎல்-இல் (Central Institute of Indian Languages (CIIL)) உள்ள தமிழ் உரை மற்றும் பேச்சு தரவுதொகுதிகளின் விவரங்கள் பின்வருமாறு அமையும்:

உரைத் தரவுத்தொகுதி:

தமிழ் சிஐஐஎல் தரவுத்தொகுதி - சொல் எண்ணிக்கை: 8068759

தமிழ் இதழ் தரவுத்தொகுதி - சொல் எண்ணிக்கை: 1805164

தமிழ் செய்தித்தாள் தரவுத்தொகுதி (வலை வலம்) - சொல் எண்ணிக்கை: 1059561

மொத்த மூல உரை சொல் எண்ணிக்கை – 10933484

அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்டது:

BIS அடையாளத் தொகுதியால் அடையாளப்படுத்தப்பட்டது - 1881646

LDC-IL அடையாளத் தொகுதியால் அடையாளப்படுத்தப்பட்டது: 58557

கூறாய்வு (chunking):

வாக்கிய எண்ணிக்கை: 20972

கூறுகளின் (chunk) எண்ணிக்கை: 160386

பேச்சுத் தரவுத்தொகுதி:

மொத்த பேச்சாளர்கள்: 453

மொத்த நேரம்: 213: 37: 27

குறிப்பு: தர சோதனை முடிந்ததும் பேச்சு தரவுத்தொகுதியின் மொத்த நேரமும் மாறும்.

12.3.1.4. சொல்சார் மூலவளங்கள் (Lexical resouces)

சுருங்ககூறினால் ஒரு சொல்சார் மூலவளம் என்பது ஒன்று அல்லது பல அகராதிகளைக் கொண்ட ஒரு தரவுத்தளமாகும். குறிப்பிடப்படும் மொழிகளின் வகையைப் பொறுத்து, சொல்சார் மூலவளம் ஒரு மொழி, இருமொழி அல்லது பன்மொழி எனத் தகுதி பெறலாம். இருமொழி மற்றும் பன்மொழி சொல்சார் மூலவளங்களில், சொற்கள் ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு இணைக்கப்படலாம் அல்லது இணைக்கப்படாமல் இருக்கலாம். இணைக்கப்படும்போது, ஒரு மொழியிலிருந்து மற்றமொழிக்கான சமன்பாடு, (இருமொழி சொல்சார் மூலவளங்களுக்கு) இருமொழிய இணைப்பு அல்லது (பன்மொழி சொல்சார் மூலவளங்களுக்கு) பன்மொழியக் குறிமானங்கள் மூலம் செய்யப்படுகிறது. ஒரே மொழியின் வெவ்வேறு அகராதிகளைக் கொண்ட ஒரு சொல்சார் மூலவளத்தை உருவாக்கவும் நிர்வகிக்கவும் முடியும்; எடுத்துக்காட்டாக,

பொதுவான சொற்களுக்கு ஒரு அகராதி மற்றும் வெவ்வேறு சிறப்பு களங்களுக்கு ஒன்று அல்லது பல அகராதிகள்.

ஒரு சொல்சார் தரவுத்தளம் (lexical database) என்பது ஒரு சொல்சார் மூலவளமாகும், இது அதன் உள்ளடக்கங்களை அணுக அனுமதிக்கவேண்டி தொடர்புடைய மென்பொருள் சூழல் தரவுத்தளத்தைக் கொண்டுள்ளது. தரவுத்தளமானது சொல்சார் தகவலுக்காக தனிப்பயன் வடிவமைக்கப்பட்டதாக (custom-designed) இருக்கலாம் அல்லது சொல்சார் தகவல் உள்ளிடப்பட்ட பொது நோக்கத்திற்கான தரவுத்தளமாக இருக்கலாம். பொதுவாக ஒரு சொல்சார் தரவுத்தளத்தில் சேமிக்கப்படும் தகவல்கள் சொல்வகைப்பாடு, சொற்களின் ஒருபொருள் பன்மொழிகள் மற்றும் வெவ்வேறு சொற்களுக்கு அல்லது சொற்களின் தொகுப்புகளுக்கு இடையிலான பொருண்மை மற்றும் ஒலியனியல் உறவுகள் ஆகியவற்றை உள்ளடக்கும்.

டைட்டி (Deparment of Electronics and Information Technology (DeitY/டைடி) நிதியளிக்கும் திராவிட சொல்வலைத் திட்டத்தின் கீழ் இந்தி சொல்வலை அடிப்படையில் தமிழ் சொல்வலை உருவாக்கப்பட்டுள்ளது. ஏறக்குறைய 30000 ஒருபொருள் பன்மொழித் தொகுதிகள் உருவாக்கப்பட்டு நான்கு முக்கிய திராவிட மொழிகளுடன் இணைக்கப்பட்டுள்ளது (Rajendran 2002, Rajendran et al 2002). திராவிடச் சொல்வலை இந்தோசொல்வலைத் திட்டத்தின் ஒரு பகுதியாகும். தகவல் பிரித்தெடுத்தல், இயந்திர மொழிபெயர்ப்பு மற்றும் பிற இயற்கைமொழி ஆய்வுப் பயன்பாடுகளுக்காக மொழியைக் கடந்து பவனிவர நிகழ்நிலை (ஆன்லைன்) சொல்சார் தரவுத்தளங்களாக சொல்வலைகள் உள்ளன.

சென்னையிலுள்ள தமிழ் இணையக் கல்விக்கழகம் நிதி உதவியால் 50000 சொற்களைக் கொண்ட தமிழ் காட்சி மூலப்பொருண்மை-சொற்களஞ்சியம் (Tamil visual onto-thesaurus) (Rajendran 2018) கோயம்புத்தூர் அமிர்தா விஸ்வ வித்யாபீடத்தால் உருவாக்கப்பட்டுள்ளது. தமிழ்ச் சொற்றொகை வகைப்பாடியல் வகைப்பாடு செய்யப்பட்டு பல்வேறு வகையான சொற்பொருண்மை அல்லது சொல் உறவுகளால் தொடர்புபடுத்தப்பட்டுள்ளது.

12.3.1.5. இலக்கணங்கள் (Grammars)

இலக்கணத்தின் வளமான வளம் தமிழில் உள்ளது. மரபுசார் கொள்கைகளை அடிப்படையாகக் கொண்டவை முதல் சாம்ஸ்கியன் ஆக்கமுறை மாற்றிலக்கண கொள்கைகள் உள்ளிட்ட நவீன கொள்கைகள் வரையிலான அனைத்து வகையான இலக்கணங்களும் தமிழின் பயன்பாட்டில். சொற்றொடர் கட்டமைப்பு இலக்கணம் (Phrase strucuture grammar), சூழல் கட்டில்லா இலக்கணம் (Contex Free Grammar (CFG)), சூழல் கட்டுண்ட இலக்கணம் (Context

Sensitive Grammar (CSG)), டாக் இலக்கணம் (Tag grammar), சொல்சார் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG)), சார்பு இலக்கணம் (Dependency Grammar), பாணினி இலக்கணம் மற்றும் தொல்காப்பியம் ஆகியவற்றை அடிப்படையாகக் கொண்ட செயல்முறைகள் தமிழைக் கணினிச்செயல் முறைக்குப் பயன்படுத்படுகின்றன. இந்தியமொழியிலிருந்து இந்தியமொழி மொழிபெயர்ப்பு (Indian Language to Indian Language Machine Translation (ILILMT) திட்டத்தில் பாணிசார் வடிவவாதம் பயன்படுத்தப்பட்டுள்ளது மற்றும் புனே சிடாக் (CDAC) ஊக்குவித்த ஆங்கிலத்திலிருந்து இந்தியமொழி மொழிபெயர்பு (Englsih to Indina Language (EILMT) திட்டத்திற்கு டேக் வடிவவாதம் பயன்படுத்தப்பட்டுள்ளது. அமிர்தா பல்கலைக்கழகம் ஆங்கிலத்திற்கும் தமிழுக்கும் இடையில் மொழிபெயர்ப்பு முறையை உருவாக்குவதற்கும் தமிழிற்கான ஆழமற்ற பகுப்பாய்வை உருவாக்குவதற்கும் சார்பு இலக்கணத்தைப் பயன்படுத்தியுள்ளது.

12.3.2. மொழி தொழில்நுட்பம்: கருவிகள், தொழில்நுட்பங்கள் மற்றும் பயன்பாடுகள்

12.3.2.1. பேச்சு அறிதல் (Speech recognition)

பேச்சிலிருந்து உரை (Speech to text (STT) அதாவது உரை பேச்சாக மாற்றுவது (Text to Speech) (TST) இயற்கைமொழ் ஆய்வின் ஒரு சவாலான பணியாகும். யெக்னநாராயணா (நயீமுல்லா கான் மற்றும் யெக்னநாராயணா 2001), ஹேமா ஏ.மூர்த்தி, ஏ.ஜி.ராமகிருஷ்ணன் மற்றும் டி.நாகராஜன் ஆகியோர் பேச்சை உரையாக மாற்றுவதற்கான அடித்தளத்தை அமைத்துள்ளனர். ஏ.ஜி.ராமகிருஹ்னன் ஒரு பெரிய சொற்றொகையை உள்ளடக்கிய தமிழ் வெளிப்பாடுகளை அறிய நரம்பியல் பின்னல் மற்றும் முற்றுநிலை மாறிகளைப் பயன்படுத்தி ஒரு தானியங்கி பேச்சு அறிதல் முறையை உருவாக்கினார். டி. நாகராஜன் DST-TIDE- நிதி உதவியுடன் "பெருமூளை வாதம் கொண்ட பேச்சாளர்களுக்கான பேச்சு-உள்ளீட்டு பேச்சு-வெளியீட்டு கருத்துப்பரிமாற்ற தொடர்பு உதவிக்கருவி' (Speech-Input Speech-Output Communicatin Aid (SISOCA) for speakers of Cerebral Palasy) என்ற மொன்பொருளை உருவாக்கினார். 'பேச்சிலிருந்து உரை'க்காக கூகிள் ஆட்மொப் (Google AdMob) ஒரு மென்பொருளைக் கொண்டுள்ளது. இது ஒலிவாங்கி மூலம் பேசப்படும் குரலை அடையாளம் கண்டு தமிழ் உரையைத் தானாகத் தட்டச்சு செய்யும். இந்த தட்டச்சு செய்த உரையை ஒருவர் சேமித்து எங்கும் பயன்படுத்தலாம். இந்த மென்பொருள் Google Chrome (பதிப்பு 25 அல்லது அதற்கு மேற்பட்டது) உலாவியில் (browser) மட்டுமே செயல்படும். கூகிள்பிளே தமிழ் பேசிலிருந்து உரை மாற்றியை வழங்குகிறது. அழகி ஆண்ட்ராய்டு பயன்பாடு-குரல் உள்ளீடு என்பது 100+ மொழிகளுக்கான பேசிலிருந்து உரை

மாறியாகும். அழகி ஆண்ட்ராய்டு பயன்பாட்டைப் பயன்படுத்தும் போது, தட்டச்சு செய்வதைத் தவிர, ஒருவர் தமிழ் உட்பட பல மொழிகளில் பேசலாம் மற்றும் ஒருவரின் உரைகளை உருவாக்கலாம். கோயம்புத்தூர் கொங்கு பொறியியல் கல்லூரியின் தங்கராஜன் தமிழிற்கான பேசிலிருந்து உரை அமைப்பின் வளர்ச்சிக்கும் பங்களித்தார் (Thankarajan et al 2008 a, 2008 b).

12.3.2.2. பேச்சு உருவாக்கம் (Speech synthesis)

பேச்சு அறிதல், அதாவது பேசிலிருந்து உரை மாற்றம் இன்னும் பாராட்டத்தக்க அளவை எட்டவில்லை என்றாலும், பேச்சு தொகுப்பு அதாவது உரையிலிருந்து பேச்சு (Text-to-speech) பாராட்டத்தக்க நிலையை எட்டியுள்ளது. ஐ.ஐ.டி. (Indian Institue of Technology (IIT/ஐ.ஐ.டி.) சென்னை தமிழின் இந்தத் தொழில்நுட்ப வளர்ச்சியில் முக்கிய பங்களிப்பை செய்துள்ளது. பெங்களூரிலுள்ள ஐ.ஐ.எஸ்.சி. (Indian Institue of Science (IISC/ஐ.ஐ.எஸ்.சி.) நிறுவனமும் பேச்சு பகுப்பாய்விற்குப் பெரும் பங்களிப்பு செய்துள்ளது. இந்தச் சூழலில் தமிழுக்கான பேச்சு தரவுத்தொகுதியை உருவாக்க மைசூரிலுள்ள சி.ஐ.ஐ.எல். நிறுவனம் மேற்கொண்ட முயற்சிகளும் பாராட்டப்பட வேண்டும்.

எழுதப்பட்ட உரையை பேச்சு வடிவமாக மாற்ற பல முயற்சிகள் உள்ளன. உரையிலிருந்து பேச்சு உருவாக்க ஒழுங்கமைப்புகளை மேம்படுத்த முன்னோடியாகச் செயல்பட்ட நிறுவனங்கள் ஐ.ஐ.டி சென்னை, சர்வதேச திராவிட மொழியியல் பள்ளி (International Institute of Dravidian Linguistics (ISDL), திருவனந்தபுரம் மற்றும் ஐ.ஐ.எஸ்.சி,, பெங்களூர், எஸ்.எஸ்.என். பொறியியல் கல்லூரி, சென்னை என்பன ஆகும். ஐ.எஸ்.டி.எல் தமிழுக்கு மிகவும் திறமையான உரையிலிருந்து பேச்சு உருவாக்க ஒழுங்கமைப்பை உருவாக்கியுள்ளது (Subramoniyam 2001). ஐ.ஐ.டி சென்னையில் இருந்த யெக்னநாராயணனால் தமிழ் உரையிலிருந்து பேச்சு உருவாக்க முயற்சி தொடங்கப்பட்டது. அவரது மாணவி ஹேமா ஏ. மூர்த்தி அவரைப் பின்தொடர்ந்து இந்த ஆராய்ச்சியில் நிறைய பங்களித்தார். அவர் தமிழிற்காக ஒரு பாராட்டத்தக்க உரையிலிருந்து பேச்சுத் ஒழுங்கமைப்பை உருவாக்கியுள்ளார். உரையிலிருந்து பேச்சு ஒழுங்கமைப்பிற்கான உள்ளீடு எப்போதும் தூய்மையான உரை அல்ல; அதில் சில தலைப்பொழுத்தன்கள், சுருக்கசொற்கள் மற்றும் தரமற்ற சொற்களாக இருக்கலாம், அவை முதலில் தொடர்புடைய முறையான தமிழ் எழுத்து வடிவத்திற்கு மாற்றப்பட வேண்டும். இந்த பணியை நிறைவேற்ற உரை இயல்பாக்கம் தொகுதி உருவாக்கப்பட்டுள்ளது. டி. நாகராஜன் (Nagarajan) 'இந்திய மொழிகளுக்கான உரையிலிருந்து பேச்சு உருவாக்கம் – ஊன உதவிக்கருவிகளுடன் ஒருங்கிணைந்த உயர்தர பேச்சிலிருந்து உரை' (Development of Text-to-Speech synthesis for

Indian Languages – High Quality TTS Integrated with Disability Aids) என்ற ஆய்த்திட்டடத்தை மேற்கொண்டார். மேலும் அவர் சென்னையிலுள்ள தமிழ் இணையக் கல்விக்கழத்தின் நிதி உதவியுடன் 'Speech enabled interactive enquiry system in Tamil' என்பதை உருவாக்கினார். மலேசிய முரசு ஒழுங்குமுறைகள் என்ற நிறுவனத்தின் நிதி உதவியுடன் எச்.எம்.எம் அடிப்படையில் 'மலேசிய தமிழர்களுக்கான உரையிலிருந்து பேச்சு ஒழுங்கமைப்பை' அவர் உருவாக்கினார். 'கட்புல ஊனம் உள்ளவர்களுக்காகப் பேச்சு உதவிக் கருவிகளை' (Speech Assistive Aids for Visually Challenged People) என்பதும் அவரால் உருவாக்கப்பட்டது. ஏ.ஜி.ராமகிருஷ்ணன் தமிழிற்கான உரையிலிருந்து பேச்சு உருவாக்கத்திற்காக ஒரு ஒழுங்கமைப்பை உருவாக்கினார் (Ramakrishnan 2007).

ஐ.ஐ.டி. சென்னையின் தலைமையில் கூட்டமைப்பு முறை திட்டத்தின் கீழ் இயந்திர வாசிப்பு உரையை மனித குரலில் வழங்கக்கூடிய ஒரு உரையிலிருந்து பேசு மென்பொருள் உருவாக்கப்பட்டது. உரையிலிருந்து பேசு மென்பொருள் மென்பொருள் பார்வைக் குறைபாடுகள் அல்லது வாசிப்பு குறைபாடுகள் உள்ளவர்கள் கணினி அல்லது மொபைல் சாதனத்தில் எழுதப்பட்ட படைப்புகளைக் கேட்க அனுமதிக்கிறது. இந்த உரையிலிருந்து பேச்சு ஒழுங்குமுறையைத் திரை வாசிப்பானுடன் ஒருங்கிணைக்கும்போது கட்புல ஊனமுள்ள பயனர்களுக்கு ஒலியுணர் இடைமுகத்தின் உதவியுடன் கணினிச் செயல்பாடுகளைப் புரிந்துகொள்ளவும் செயல்படுத்தவும் உதவும். இவ்வாறு திரை வாசிப்பான்ன் ஒருங்கிணைந்த உரையிலிருந்து பேசு ஒழுங்கமைப்பு ஒரு திறனுள்ள உதவி தொழில்நுட்பமாக இருக்கும். இது பார்வை-சவால் உள்ள பிரிவினருக்கும் சமூகத்தின் கற்றல் குறைபாடுள்ள பிரிவினருக்கும் ஐ.சி.டி (Information and communications technology (ICT)) மற்றும் அறிவுப் பகிர்வின் நன்மைகளைப் பயன்படுத்த உதவும்.

12.3.2.3. இலக்கண பகுப்பாய்வு (Grammatical analysis)

இயற்கையான மொழியின் முக்கியமான இலக்கணப் பகுப்பாய்வு ஒலியனியல் பகுப்பாய்வு, உருபனியல் பகுப்பாய்வு, தொடரியல் பகுப்பாய்வு ஆகியவற்றை உள்ளடக்கும்.

தமிழ் ஒப்பீட்டளவில் சொல்வரிசைக் கட்டுப்பாடு இல்லாத, வினை இறுதி மற்றும் திரிபுறும் சொற்கள் கொண்ட மொழி ஆகும். படர்க்கை இடப் மாற்றுப் பெயர்கள் (பதிலீடு பெயர்கள்) ஒற்றை-பன்மை வேறுபாடு மற்றும் ஆண்பால்-பெண்பால் வேறுபாட்டைக் காட்டுகின்றன. தமிழ் மொழியில், வினைச்சொல்லைப் பொறுத்து ஒரு வாக்கியத்தின் பங்கேற்பாளர்கள் (எழுவாய், செயப்படுபொருள், மறைமுகச் செயப்படுபொருள், கருவி மற்றும்

இருப்பிடம் போன்றவை) அவை ஏற்கும் வேற்றுமை உருபுகளால் அடையாளம் காணப்படும். தமிழ் ஒரு ஒட்டுநிலை மொழி என்பதால், சொற்கள் பல உருபன்களை இணைப்பதன் மூலம் உருவாகின்றன. ஒரு தமிழ் சொல் என்பது பிற இலக்கணச் சேர்க்கைகளுடன் இணைந்த ஒரு வேரைக் கொண்ட ஒரு கலவையாகும். தமிழ்ச் சொற்களின் நீளம், சிக்கலான தன்மை மற்றும் வகையைப் பொருட்படுத்தாமல், அடுத்தடுத்த சேர்க்கைகளைக் கவனமாக அகற்றுவதன் மூலம் வேர்களை ஓர்சை நிலை வரை அடையாளம் காணலாம். பாரம்பரியமாக, ஒரு தமிழ் சொல் அதிக அளவில் ஆறு பகுதிகளாகப் பிரிக்கப்பட்டுள்ளது, அதாவது பகுதி (முதன்மைப் பகுதி), சந்தி (சந்தி), விகாரம் (மாறுபாடு), இடைநிலை (நடுப் பகுதி), சாரியை (ஒலி இசைவி) மற்றும் விகுதி (முற்றுருபு) . எடுத்துக்காட்டாக, நடந்தனன் என்ற சொல் பின்வரும் உருபன்களால் ஆனது: நட+ (த்) ந் + த் + அன் + அன். நடுப் பகுதியும் மற்றும் விகுதியும் முதன்மைப் பகுதிக்கு இலக்கணம் சேர்க்கும். நடுப் பகுதி காலத்தையும் விகுதி பாலினத்தையும் குறிக்கிறது. வழக்கமாக, முதன்மைப் பகுதி என்பது அதன் பொருளுக்குப் பொறுப்பான சொல்லின் முக்கிய பகுதியாகும்.

12.3.2.3.1. டோகன்களாகப் பிரித்தல் (Tokenization)

டோக்கன்களாகப் பிரித்தல் (Tokenization) என்பது கொடுக்கப்பட்ட உரையை டோக்கன்கள் (tokens) எனப்படும் அலகுகளாகப் பிரிக்கும் செயல்முறையாகும். டோக்கன் சொல்லாகவோ எண்ணாகவோ நிறுத்தற்குறியாகவோ இருக்கலாம். சொல் அல்லது பிற அலகுகளின் எல்லைகளைக் கண்டறிவதன் மூலம் டோக்கன்களாகப் பிரித்தல் இந்த பணியை செய்கிறது. ஒரு சொல்லின் இறுதியும் அடுத்த சொல்லின் தொடக்கமும் சொல் எல்லை என்று அழைக்கப்படுகிறது. டோக்கனைசேஷன் என்பதை நடைமுறைப் பயன்பாட்டிற்காகச் சொல் பிரித்தல் என எளிமையாகப் புரிந்துகொள்ளலம். எடுத்துக்காட்டாக, டோக்கன் பிரிப்பான் *'இராமன் ஒரு பழம் தின்றான்.*' என்ற வாக்கியதை *இராமன், ஒரு, பழம், தின்றான்*, . (நிறுத்தற்குறி) என டோக்கன்களாகப் பிரிக்கும். டோக்கன்களாகப் பிரிக்கும் செயல்பாட்டைச் செய்யும் கணினி ஒழுங்கமைப்பு டோக்கன் பிரிப்பான் (Tokenizer) எனப்படும்.

டோக்கனைசேஷன் தொடர்ச்சியான உரையின் ஒரு பகுதியைத் தனித்தனிச் சொற்களாகப் பிரிக்கும். ஆங்கிலம் போன்ற மொழியைப் பொறுத்தவரை இது மிகவும் சுலபமானது; ஏனெனில் சொற்கள் பொதுவாக இடைவெளிகளால் பிரிக்கப்படுகின்றன. இருப்பினும், தமிழ் போன்ற ஒட்டுநிலை மொழிகள் அத்தகைய முறையில் சொல் எல்லைகளைக் குறிக்கவில்லை; மேலும் அந்த மொழிகளில் உரையைப் பிரிப்பது ஒரு முக்கியமான பணியாகும். அதற்கு அம்மொழியில் உள்ள சொற்றொகை மற்றும் சொற்களின் உருபனியல் பற்றிய அறிவு தேவைப்படும். சில நேரங்களில்

இந்தச் செயல்முறை தரவு அகழ்ந்தெடுப்பில் (data mining) சொற்பை (Bag of Words (BOW) உருவாக்கம் மற்றும் சொற்பொருள் மயக்கநீக்கம் (Word Sense Disambiguation (WSD) போன்ற நிகழ்வுகளிலும் பயன்படுத்தப்படுகிறது.

தமிழை நோக்கிய அனைத்து எந்திரமொழிபெயர்பு அமைப்புகளும் டோக்கனைசேஷனில் தொடங்கப்பட வேண்டும்.

12.3.2.3.2. பகுதி ஆய்தல் (Stemming)

மொழியியல் உருபனியல் மற்றும் தகவல் மீட்டெடுப்பில், பகுதியைப் பிரித்தல் என்பது திரிபுற்ற (அல்லது சில நேரங்களில் ஆக்கப்பட்ட) சொற்களை அவற்றின் சொல் பகுதியாக அல்லது அடிப்படை அல்லது வேர் வடிவமாகக் குறைக்கும் செயல்முறையாகும். பகுதி என்பது பொதுவாக எழுதப்பட்ட ஒரு சொல் வடிவம் ஆகும். பகுதி சொல்லின் உருபனியல் வேருக்கு ஒத்ததாக இருக்க வேண்டியதில்லை; இந்த பகுதி ஒரு சரியான வேராக இல்லாவிட்டாலும், தொடர்புடைய சொற்கள் ஒரே பகுதியுடன் தொடர்புடையதாக இருப்பது போதுமானது. எடுத்துக்காட்டாக, *வா, வந்தான், வருகிறான், வருவான்* என்ற சொற்களை ஆய்ந்து *வா, வ, வரு* என்ற ஒரே பொருளைத் தரும் பகுதிகளை அடையாளம் காண்பது பகுதி ஆய்தல் எனப்படும். சொற்களை உருபன்களாகப் பிரித்து பகுதியை அடையாளம் காணும் கணினி ஒழுங்குமுறை பகுதி ஆய்வி (stemmer) எனப்படும்.

பகுதி ஆய்தல் அல்லது பிரித்தலுக்கான வழிமுறைகள் 1960களில் இருந்து கணினி அறிவியலில் ஆய்வு செய்யப்பட்டுள்ளன. பல தேடுபொறிகள் ஒரே மாதிரியான பகுதி சொற்களை ஒரு வகையான வினவல் விரிவாக்கமாகக் கருதுகின்றன, இது வேறுபட்டவைகளை ஒருங்கிணக்கும் (conflation) ஒரு செயல்முறை ஆகும். ஒரு சொல்லின் பகுதியைப் பிரிக்கும்/ஆயும் ஒரு கணினி நிரல் அல்லது துணைநிரல், ஒரு பகுதி ஆய்வு நிரல், பகுதி ஆய்வு வழிமுறை என்று அழைக்கப்படலாம்.

தமிழிற்காக ஏராளமான பகுதி ஆய்வி ஒழுங்கமைப்புகள் உருவாக்கப் (Thkarasu and Manavalan 2013). அவை தரவுத்தொகுதி அல்லது உரைகளை டோக்கன்களாகப் பிரிக்கப்பட்ட சொற்களிலிருந்து பகுதியைப் பிரிக்கப்படுப் பயன்படுகின்றன.

12.3.2.3.3. சொல்லன் ஆய்தல் (Lemmatization)

மொழியியலில் சொல்லன் ஆய்தல் (Lemmatization) என்பது ஒரு சொல்லின் திரிபுற்ற வடிவங்களை ஒன்றிணைக்கும் செயல்முறையாகும்; இதன்படி திரிபுற்ற வடிவங்களை ஒரு பொதுவான வடிவத்திலிருந்து விளைந்ததாகக் கருத் இயலும். இது சொல்லின் சொல்லன் அல்லது

அகராதி வடிவத்தால் அடையாளம் காணப்படுகிறது. கணினி மொழியியலில் சொல்லன் ஆய்தல் என்பது ஒரு சொல்லின் சொல்லனை அதன் கருதப்பட்ட பொருண்மையின் அடிப்படையில் தீர்மானிக்கும் வழிமுறையாகும். பகுதி ஆய்வு போலல்லாமல் சொல்லன் ஆய்வு, ஒரு வாக்கியத்தில் மட்டுமன்றி அண்டை வாக்கியங்கள் அல்லது ஒரு முழு ஆவணம் போன்ற அந்த வாக்கியத்தைச் சுற்றியுள்ள பெரிய சூழலுக்குள்ளும் ஒரு சொல்லின் கருத்தப்பட்ட சொல்வகைப்பாடு மற்றும் பொருண்மைச் சரியாக அடையாளம் காணப்படுவதைச் சொல்லன் ஆய்தல் சார்ந்துள்ளது. இதன் விளைவாக, திறமையான சொல்லன் ஆய்வு வழிமுறைகளை உருவாக்குவது ஆராய்ச்சியின் திறந்த பகுதி ஆகும். எடுத்துக்காட்டாக வா, *வந்தான், வருகிறான், வருவான்* என்பதை உருபன்களாகப் பிரிக்கையில் *வா, வ, வரு* என்ற பகுதிகள் விளையும். இப்பகுதிகள் எல்லாம் ஒரே பொருளை உணர்த்துவதன் காரணமாக அவை *வா* என்ற சொல்லனின் (அகராதிப் பதிவின்) வேறுபட்ட வடிவங்களாகக் கருதப்படவேண்டும். சொல்லன் ஆய்வை செயல்படுத்தும் கணினி ஒழுங்கமைப்பு சொல்லன் ஆய்வி என்று அழைக்கப்படுகிறது.

அமிர்தா பல்கலைக்கழகம், ஏ.யூ.கே.பி.சி.ஆர்.சி. மற்றும் அண்ணா பல்கலைக்கழகம் ஆகியநிறுவனங்களாலும் சில தனிநபர்களாலும் சொல்லன் ஆய்விகள் உருவாக்கப்பட்டுள்ளன.

12.3.2.3.4. உருபனியல் பகுப்பாய்வு (Morphological analysis)

உருபனியல் பகுப்பாய்வு என்பது சொற்களைத் தனித்தனி உருபன்களாகப் பிரித்து உருபன்களின் வகுப்பை அடையாளம் காணும் செயல்முறையாகும். இந்தச் செயல்பாட்டின் கடினம் கருதப்படும் மொழியின் உருபனியல் (அதாவது சொற்களின் அமைப்பு) சிக்கலைப் பொறுத்தது. ஆங்கிலத்தில் மிகவும் எளிமையான உருபனியல் உள்ளது, குறிப்பாக திரிபுசார் உருபனியல்; எனவே இந்தச் செயல்பாட்டை முற்றிலுமாகப் புறக்கணித்து, ஒரு சொல்லின் சாத்தியமான அனைத்து வடிவங்களையும் (எ.கா.open, opens, opened, opening) தனித்தனிச் சொற்களாக கணிக்க இயலும் எளிய மாதிரி போதுமானது. தமிழ் மிகவும் மேம்பட்ட ஒட்டுநிலை ஒருங்கிணைந்த இந்திய மொழியாகும்; எனவே, அத்தகைய அணுகுமுறை சாத்தியமில்லை; ஏனெனில் ஒவ்வொரு அகராதி பதிவும் ஆயிரக்கணக்கான சொல் வடிவங்களைக் கொண்டிருக்க வேண்டிவரும். எனவே தமிழுக்கு உருபனியல் ஆய்வி மிகத் தேவையானதாகும். உபனியல் பகுப்பாய்வு செய்யும் கணினி ஒழுங்கமைப்பு உருபனியல் பகுப்பாய்வி என்று அழைக்கப்படுகிறது. இப்பகுப்பாய்வி *வந்தான்* என்ற சொல்லை வ+ந்த்+ஆன் என உருபன்களாகப் பகுப்பாய்வு செய்யும். மட்டுமன்றி *வ* என்பது ‘come’ என்று பொருள்படும்

என்றும், *த்த்* என்பது இறந்தகாலம் உணர்த்தும் என்றும் *ஆன்* என்பது ஆண்பால் ஒருமையைக் குறிக்கும் என்ற தகவல்களையும் தரும்.

இயந்திர மொழிபெயர்ப்பு (Machine Translation (MT)) உள்ளிட்ட பல என்எல்பி பயன்பாடுகளுக்கு உருபனியல் பகுப்பாய்வி உருவாக்குவது முக்கியமானது. தமிழுக்கான தற்போதைய உருபனியல் பகுப்பாய்விகள் பிரிப்பதை முக்கியமாகச் சாத்தியமான அனைத்து உருபுகளையும் உருவாக்கச் செயல்படுத்தும்; பின்னர் இலக்கண விதிகள் அல்லது டேக்கிங் பொருந்தாத தன்மையைப் பயன்படுத்தி பின் ஆய்வின் போது சரியானவைகளைத் தேர்ந்தெடுக்கும்.

ஐ.ஐ.டி. கான்பூரில் மேற்கொள்ளப்பட்ட அனுசாரக திட்டத்தின் கீழ் தமிழுக்கான முதல் உருபனியல் பகுப்பாய்வி உருவாக்கப்பட்டது. அனுசாரகா திட்டத்தில் உருபனியல் பகுப்பாய்வுத் தொகுதிக்காக உருவாக்கப்பட்ட சட்டகத்தின் அடிப்படையில் ராஜேந்திரன் தமிழிற்கான முதல் உருபனியல் பகுப்பாய்வியை உருவாக்கினார். பின்னர் அதை AUKBCRC மேம்படுத்தியது. தமிழுக்கான உருபனியல் பகுப்பாய்வியை உருவாக்குவதில் பின்வரும் நிறுவனங்கள் குறிப்பிடத் தகுந்தன: இந்திய மொழிகளின் மைய நிறுவனம் (Ganesan 1994; Ganesan and Francis Ekka (1994), ஏ.யூ.கே.பி.சி.ஆர்.சி (Viswanathan 2000; Viswanathan et al 2003; Vijay Sundar Ram et al 2010; Marimuthu et al 2013), அண்ணா பல்கலைக்கழகம் (Anandan et al 2002), அமிர்தா விஷ்வ வித்யபீடம், கோயம்பத்தூர் (Dhanlakshmi et al 2009a. 2009b; Menon 2009; Anand Kumar et al 2009; Anand Kumar et al 2010a; Anand Kumar et al 2010b), தமிழ்ப் பல்கலைக்கழகம் (Rajendran 1999; Rajendran et al 2003, Rajendran 2006), மத்தியப் பல்கலைக்கழகம், ஹைதராபாத் (Arulmozi 1998; Winston Cruz 2002; Ramaswami 2003), சென்னைப் பல்கலைக்கழகம் (Deivasundaram and Gopal 2003), அண்ணாமலைப் பல்கலைக்கழகம் (Ganesan 2003). சில தனிப்பட்ட முயற்சிகளும் மேற்கொள்ளப்பட்டன. எடுத்துக்காட்டாக, தெய்வசுந்திரத்தின் தமிழ்ச் சொல்லாய்வி/சொல்லாளர் மற்றும் கணேசனின் உருபனியல் பகுப்பாய்வி இவற்றைக் கூறலாம். அகிலன் மற்றும் பங்காளிகள் (Ahilan et al 2015) விதிமுறை அணுகுமுறை அடிப்படையில் செம்மொழி உரைக்கு ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கியுள்ளனர். வெளிநாட்டிலிருந்தும் சில குறிப்பிடத்தக்க பங்களிப்புகள் செய்யப்பட்டுள்ளன (Vasu Ranganathan 1997, Loganathan et al Lushanthan 2014; Mokanarangan et al 2016, Ranganathan 2016 b).

பல்வேறு வகையான உருனியல் பகுப்பாய்விகள் உருவாக்கப்பட்டுள்ளன. விதி அடிப்படையிலான அணுகுமுறை, வரிசை புலக்குறிப்புசெய்யும் அணுகுமுறை மற்றும் இயந்திர கற்றல் அணுகுமுறை போன்ற வெவ்வேறு அணுகுமுறைகள் கையாளப்பட்டுள்ளன.

12.3.2.3.5. உருபனியல் உருவாக்கம் (Morphological Generation)

உருபனியல் உருவாக்கம் உருபனியல் ஆய்வின் மறுதலையாகும். ஒரு சொல்லின் உள்ளமைப்பு உருப்படுத்தத்தை புற அமைப்பு வடிவமாக மாற்றும் செயல்பாடு உருபனியல் உருவாக்கம் எனப்படும். உருபனியல் உருவாக்கத்தை செயல்படுத்தும் கணினி ஒழுங்கமைப்பு உருபனியல் உருவாக்கி எனப்படும். எடுத்துக்காட்டாக உருபனியல் உருவாக்கி *வ, ந்த், ஆன்* என்ற உருபன்களை இணைத்து *வந்தான்* என்ற சொல் வடிவை உருவாக்கும்.

பெரும்பாலான இயந்திர மொழிபெயர்ப்பு அமைப்பில் இறுதி தொகுதி உருபனியல் உருவாக்கி ஆகும். ஒரு குறிப்பிட்ட இலக்கண வகைக்கு ஒரு சொல் வடிவத்தை உருவாக்க, அதனுடன் தொடர்புடைய ஒட்டுகள் வேர்ச் சொல்லுடன் இணைக்கப்பட வேண்டும். உருபனியல் உருவாக்கி முந்தைய தொகுதியிலிருந்து அதன் உள்ளீட்டை எடுக்கிறது. உள்ளீடு சொல்லின் வேரையும் அதன் இலக்கணக் கூறுகளையும் கொண்டிருக்கும். உருபனியல் உருவாக்கி பின்னர் இலக்கு மொழியின் உருபனியல் விதிகளின் அடிப்படையில் வேர்ச்சொல்லுடன் தொடர்புடைய இலக்கணக்கூறுகளை வெளிப்படுத்தும் ஒட்டுக்களை இணைத்து இலக்கு மொழிச் சொல் வடிவத்தை உருவாக்கி வெளியீடு செய்யும். இவ்வாறு உருவாக்கப்படும் சொற்கள் முழுமையான இலக்கு மொழி வாக்கியத்தை உருவாக்குவதற்கு வேண்டி ஒன்றிணைக்கப்படுகின்றன. தமிழைக் குறிவைக்கும் இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்புகள் உருபனியல் உருவாக்கிகளை அல்லது சொல் உருவாக்கிகளை உருவாக்கின. ILILMT cosortitum project மற்றும் EILMT consortitum project ஆகியவை தமிழிற்கான உருபனியல் உருவாக்கியின் ஆகத்திற்குப் பங்களிப்பு செய்தன. சி.ஐ.ஐ.எல்., ஹைதராபாத் மத்திய பல்கலைக்கழகம், (Winston Cruz 2002; Ramasway 2000) ஏ.யூ.கே.பி.சி.ஆர்.சி (Menaka et al 2018), அமிர்தா விஷ்வ வித்யபீடம் (Manon 2009; Anandkumar et al 2010) மற்றும் அண்ணா பல்கலைக்கழகம் (Anandan et al 2001) அவற்றின் இயற்கை மொழி ஆய்வுத் திட்டங்களுக்கான தேவைகளைப் பூர்த்தி செய்ய வேண்டி தமிழுக்கான உருபனியல் உருவாக்கியை ஆக்கின. தனிநபர்களும் வணிக நிறுவனங்களும் கூட தமிழ் உருபனியல் உருவாக்கியை ஆக்க முயன்றன. எடுத்துக்காட்டாக அல்டிமேட் சாப்ட்வேர் சொல்யூஷன், திண்டுக்கல் என்ற நிறுவனத்தைச் சார்ந்த துரைபாண்டி உருவாக்கிய 'தமிழ் வினை

வடிவங்களுக்கான உருபனியல் உருவாக்கி மற்றும் பகுப்பாய்வு இயந்திரதிரம்' என்ற மென்பொருளைக் கூறலாம் (துரைபாண்டி 2006).

12.3.2.3.6. சொல்வகைப்பாடு அடையாளப்படுத்தல்

சொல்வகை அடையாளப்படுத்தல் (Parts of Speech tagging (POS tagging)) செயல்பாடு ஒரு வாக்கியத்தைக் கொடுத்தால், ஒவ்வொரு சொல்லின் சொல்வகைப்பாட்டையும் தீர்மானிக்கும். சொல்வகையை அடையாளப்படுத்தும் கணினி ஒழுங்கமைப்பு சொல்வகை அடையாளப்படுத்தி (Parts of Speech tagger (POS tagger)) எனப்படும். எடுத்துகாட்டாக *இராமன் மாம்பழம் தின்றான்* என்ற வாக்கியம் தரப்படுகையில் சொல்வகை அடையாளப்படுத்தி *இராமன்* என்பதை பெயர் எனவும் *மாம்பழம்* என்பதை பெயர் எனவும் *தின்றான்* என்பதை வினை எனவும் அடையாளப்படுத்தும். பல சொற்கள், குறிப்பாகப் பொதுவானவை, பல சொல்வகைப்பாடுகளாகச் செயல்படக்கூடும். சில மொழிகள் பிற மொழிகளை விட அதிக அளவில் சொல்வகைப்பாட்டு மயக்கம் காட்டும். குறிப்பாக ஆங்கிலம் போன்ற குறைந்த அளவில் திரிபுறும் மொழிகள் இத்தகைய சொல்வகைப்பாட்டு மயக்கத்திற்கு ஆளாகின்றன. சீன மொழி இத்தகைய சொல்வகைப்பாட்டு மயக்கத்திற்கு ஆளாகின்றது; ஏனெனில் இது பேசப்படுகையில் போது ஒரு சுரம்சார் மொழி (tonal language) ஆகும். தமிழ் திரிபு அடிப்படையில் செழுமையான மொழி. சொல்வகை அடையாளப்படுத்தல் இலக்கண வகைப்பாடு மயக்கதை அகற்றலாம். (இரு சொல்வகைப்பாடுகளைச் சார்ந்த ஒருசொல் போலிகள் ஒரே வடிவில் இருப்பதால் சொல்வகைப்பாட்டு மயக்கம் ஏற்படும்).

ஒவ்வொரு சொல்லின் சொல்வகைப்பாட்டை தீர்மானிக்கும் ஒழுங்கமைப்பு சொல்வகைப்பாட்டு அடையாளப்படுத்தி (POS tagger) என்று அழைக்கப்படுகிறது. தமிழுக்கான சொல்வகைப்பாட்டு அடையாளத் தொகுதிகள் (POS tag set) பல நிறுவனங்களாலும் தனிநபர்களாலும் உருவாக்கப்பட்டன. இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சித் திட்டம் (Technological Development of Indian Languages (TDIL/டி.டி.ஐ.எல்) தமிழ் உள்ளிட்ட எல்லா இந்திய மொழிகளுக்காக பி.ஐ.எஸ்./பிஸ் (Bureau of Indian Standards (BIS)) சொல்வகைப்பாட்டு அடையாளத்தொகுதியை அறிமுகப்படுத்தியது. இந்திய மொழிகளின் தரவுத்தொகுதிகள் முன்முயற்சி (Indian Languages Corpora Initiative (ILCI/ஐ.எல்.சி.ஐ.) திட்டம் உட்பட அனைத்து இந்திய அரசின் நிதியுதவி திட்டங்களும் இதைப் பின்பற்றுகின்றது. மைக்ரோசாஃப்ட், இந்தியா 'இந்திய மொழிகளுக்கான பொதுவான சொல்வகைப்பாடு அடையாளத்தொகுதி' சட்டகத்தை உருவாக்க முயற்சித்தது (Sanakaran Baskaran et al 2008). இது ஒரு படிநிலை

அடையாளத்தொகுதி (hierarchical tagset) ஆகும். தமிழுக்குக் கிடைக்கும் மற்றொரு வகை அடையாளத்தொகுதி தட்டையான அடையாளத்தொகுதி (flat tagset) ஆகும்.

இந்தியமொழிகளின் மையநிறுவனம் தமிழ் தரவுத்தொகுதியைச் சொல்வகைப்பாட்டிற்கு அடையாளப் படுத்த வேண்டி சொல்வகைப்பாட்டு அடையாளத் தொகுதியை உருவாக்கியது. அமிர்தா பல்கலைக்கழகம் அதன் இயற்கைமொழி ஆய்வுத்திட்டங்களுக்காக ஒரு சொல்வகை அடையாளத்தொகுதியை உருவாக்கிப் பயன்படுத்தியது (Dhanlakshmi et al 2009). வெளிநாட்டு நிறுவனங்களும் தனிநபர்களும் சொல்வகைப்பாட்டு அடையாளத் தொகுதி உருவாக்கத்திற்கு குறிப்பிடத்தக்க பங்களிப்பு செய்தன (Ranganathan 2001; Thyaparan et al 2018). சொல்வகைப்பாட்டு அடையாளப்படுத்தலை மேம்படுத்துவதற்கான முயற்சிகளும் மேற்கொள்ளப்பட்டன (Selvam et al 2009). சொல்வகைப்பாட்டு அடையாளப்படுத்தலுக்கு இரண்டு முக்கிய வகையான அணுகுமுறைகள் பயன்படுத்தப்படுகின்றன: விதி அடிப்படையிலான அணுகுமுறை மற்றும் இயந்திர கற்றல் அணுகுமுறை (ஆனந்தகுமார் மற்றும் பலர், தனலட்சுமி மற்றும் பலர்). ராஜேந்திரன் (2007) தமிழில் சொல்வகைப்பாட்டு அடையாளப்படுத்தலின் சிக்கலைப் பற்றி விவாதிக்கிறார்.

12.3.2.3.7. தொடரியல் பகுப்பாய்வு (Syntactic parsing)

தொடரியல் பகுபாய்வு கொடுக்கப்பட்ட வாக்கியத்தின் படிநிலை அமைப்பைத் (இலக்கண பகுப்பாய்வு) தீர்மானிக்கிறது. இயற்கையான மொழிகளுக்கான இலக்கணம் தெளிவற்றது மற்றும் வழக்கமான வாக்கியங்கள் பல சாத்தியமான பகுப்பாய்வுகளைக் கொண்டுள்ளன. உண்மையில், ஒருவேளை ஆச்சரியப்படும் விதமாக, ஒரு பொதுவான வாக்கியத்திற்கு ஆயிரக்கணக்கான சாத்தியமான பகுப்பாய்வுகள் இருக்கலாம் (அவற்றில் பெரும்பாலானவை ஒரு மனிதனுக்கு முற்றிலும் முட்டாள்தனமாகத் தோன்றும்). பகுப்பாய்வில் இரண்டு முதன்மையான வகைகள் உள்ளன: அண்மை உறுப்புப் பகுப்பாய்வு மற்றும் சார்புப் பகுப்பாய்வு. அண்மை உறுப்புப் பகுப்பாய்வு (சூழல் கட்டில்லா இலக்கணம் போன்று) வாக்கியங்களை அண்மை உறுப்புகளாகப் (தொடர்களாகவும் சொற்களாகவும்) பகுத்து ஒரு படிநிலை அமைப்பை உருவாக்குவதில் கவனம் செலுத்துகிறது. சார்புப் பாகுப்பாய்வு ஒரு வாக்கியத்தில் உள்ள சொற்களுக்கு அல்லது தொடர்களுக்கு இடையிலான உறவில் (பயனிலை, எழுவாய், செயப்படுபொருள், அயல்செயப்படுபொருள்) கவனம் செலுத்துகிறது. வாக்கியங்களைப் பகுப்பாய்வு செய்யும் கணினி ஒழுங்குமுறை தொடரியல் பகுப்பாய்வி ஆகும் (syntactic parser)

AUKBCRC தமிழிற்கான ஒரு தொடரியல் பகுப்பாய்வு உருப்படுத்தத்தை உருவாக்கியது (Kumar Shanmugam 2004). தமிழுக்கு பல்வேறு வகையான தொடரியல் பகுப்பாய்விகள் முயற்சிக்கப்பட்டுள்ளன. அவை சொல்சார் மற்றும் புள்ளியியல் பகுப்பாய்வு, சார்புப் பகுப்பாய்வு, டாக் வடிவவாதம் அடிப்படையிலான பகுப்பாய்வு (Menon et al 2014), ஆழமில்லாப் பகுப்பாய்வு (Arirathnam et al 2015), தொடரமைப்புக் கலப்பின மொழி மாதிரியைப் பயன்படுத்தி தொடரமைப்புப் பகுப்பாய்வு (Selvan et al 2009), சார்பு மாதிரியைப் பயன்படுத்தி தொடரமைப்புப் பகுப்பாய்வு (Selvam et al 2008), பென் கிளையமைப்பு வங்கி அடிப்படையிலான தொடரியல் பகுப்பாய்வு (Antony et al 2010), விதி அடிப்படையிலான மற்றும் தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகளைப் பயன்படுத்தி சார்புப் பகுப்பாய்வு (Ramaswamy et al 2011) என்பனவாகும். வெளிநாட்டிலிருந்து வந்த பங்களிப்புகளில் வாசு ரங்கநாதனின் தமிழுக்கான தொடரியல் பகுப்பாய்வி குறிப்பிடத்தகுந்த ஒன்றாகும் (Ranganathan 2016b).

12.3.2.3.8. கூறாய்வு (chunking)

முழுமையான தொடரியல் பகுப்பாய்விற்குப் பதிலாக கூறாய்வு (chunking) இயற்கை மொழி ஆய்வில் பரிந்துரை செய்யப்படுகின்றது. முழுமையான தொடரியல் பகுப்பாய்வு அதிக காலம் எடுக்கும் மற்றும் பொருளாதாரீதியாக உகந்தது அல்ல. இயந்திர மொழிபெயர்ப்பு போன்ற இயற்கை மொழி ஆய்வுச் செயல்பாடுகளின் வாக்கியங்களைத் தகுந்த கூறுகளாகப் (chunks) பகுத்தால் போதுமானது. இத்தகைய செயல்பாட்டைச் செய்யும் கணினி ஒழுங்குமுறை கூறாய்வி (chunker) எனப்படும்.

கூறாய்வி ஒரு வாக்கியத்தை மேலுறல் செய்யாத தொடர்களாகப் பிரித்து அவை ஒவ்வொன்றிற்கும் ஒரு புலக்குறிப்பு (label) தரும். கூறாக்கிகள் அவற்றின் துல்லியமான வெளியீடு மற்றும் கூறு வரையறுக்கப்பட்டும் விதம் இவற்றில் வேறுபடுகிறது. பலர் கூறுகளாகப் பிரிப்பதை விட அதிகமாகச் செயல்படுவர். பிறர் பெயர்த்தொடர்களை மட்டும் கண்டுபிடிப்பார்கள். சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கும் முழு பகுப்பாய்வுக்கும் இடையில் கூறாய்வு வரும். தனிப்பட்ட கூறுகளின் கட்டமைப்பை விளக்குவது மிகவும் எளிதானது; அதே நேரத்தில் கூறுகளுக்கு இடையிலான உறவுகளை விளக்குவது கடினமானது மற்றும் தனிப்பட்ட சொல்சார் பண்புகளைச் சார்ந்துள்ளது. கூறாய்வி வாக்கியத்தைக் கூறுகளாக ஆய்ந்து கூறுகளை அவற்றின் செயல்பாடு அடிப்படையில் அடையாளப் படுத்துகின்றது. ஒரு எடுத்துக்காட்டான கூறு ஒரு பொருளக்கச் சொல்லையும் அதைச் சுற்றி ஒரு குழுமச் செயல்பாட்டுச் சொற்களையும்

கொண்டிருக்கும். கூறுகள் பொதுவாக மறுநிகழ்வு இல்லாத ஒன்றோடொன்று தொடர்புடைய சொற்களாகக் கருதப்படுகின்றன. தமிழ் ஒரு ஒட்டுநிலை மொழியாக இருப்பதால் ஒரு சிக்கலான உருபனியல் மற்றும் தொடரியல் கட்டமைப்பைக் கொண்டுள்ளது. இது ஒப்பீட்டளவில் சொல் வரிசைக் கட்டுப்பாடில்லாத மொழி; ஆனால் சொற்றொடர் மற்றும் எச்சத்தொடர் கட்டுமானத்தில் இது ஒரு நிலையான சொல் சொல்வரிசை உள்ள மொழியாகச் செயல்படுகிறது. எனவே சொல்வகைப்பாடு அடையாளப்படுத்தும் செயல்முறையுடன் ஒப்பிடும்போது தமிழில் கூறாய்வுச் செயல்முறை குறைவான சிக்கலானது.

AUKBCRC தமிழுக்கு ஒரு பெயர்ச்சொல் கூறாய்வியை உருவாக்கியுள்ளது (Sobha & Vijay Sundar Ram 2006). தமிழில் கூறாய்வு குறித்த விரிவான செயல்பாடு அமிர்தா விஷ்வ விதியப்பீடம் நிறுவனத்தால் மேற்கொள்ளப்பட்டுள்ளது Dhanalakshmi (2010; Dhanalakshmi et al 2019).

12.3.2.3.9. ஆழமற்ற பகுப்பாய்வு

தமிழில் ஆழமற்ற பகுப்பாய்வு குறித்த ஒரு முன்னோடியான ஆராய்ச்சி அமிர்தா விஷ்வ விதியப்பீடம் நிறுவனத்தால் மேற்கொள்ளப்பட்டுள்ளது (Dhanalakshmi 2010) ஆழமற்ற பகுப்பாய்வு பின்வரும் நான்கு தொகுதிகளைக் (modules) கொண்டுள்ளது: (i) சொல்வகைப்பாட்டு அடையாளங்களால் வாக்கியங்களில் வரும் சொற்களை அடையாளப்படுத்தும் சொல்வகை அடையாளப்படுத்தி (ii) ஆங்கில அர்த்தங்களைத் தந்து சொற்களை உருபன்களாகப் பகுப்பாய்வு செய்யும் உருபனியல் பகுப்பாய்வி, (iii) வாக்கியத்தை மேலுறல் செய்யாத (பெயர்த்தொடர், வினைத்தொடர் முதலிய) தொடர்களாகப் பகுத்து புலக்குறிப்பு செய்யும் கூறாய்வி (iv) உறவைக் கண்டுபிடிப்பதற்கான சார்பு பகுப்பாய்வி: ஒரு வாக்கியத்தில் வரும் சொற்களின் சொல்வகைப்பாடு மற்றும் கூறாய்வுக் கூறுகள் தரப்படுகையில், இது வினைச்சொல்லுடன் கூறாய்வுக் கூறுகள் கொண்டுள்ள உறவுகளை (எழுவாய், செயப்படுபொருள், அயற் செயப்படுபொருள் போன்றவற்றை) தீர்மானிக்கிறது. தமிழுக்கான ஆழமற்ற பகுப்பாய்வி இயந்திர கற்றல் அணுகுமுறையைப் பயன்படுத்தி சொல்வகை அடையாளப்படுத்தலையும் கூறாய்வு செய்வதையும் கூறுகளுக்கு இடையிலான உறவுகளை நிறுவதையு செய்கின்றது. அதற்காக உருவாக்கப்பட்ட தொகுதிகள் நம்பகமான முடிவுகளைக் காட்டுகின்றன. 10,000 சொற்களின் நிகழ்நேர சொற்றொகுதியில் மதிப்பீடு சோதனை செய்யும் போது உருபனியல் ஆய்வி 95% துல்லியத்தையும் எஸ்.வி.எம்.டூலைப் பயன்படுத்தும் சொல்வகை அடையாளப்படுத்தி 94.12% துல்லியத்தை தந்தது. மால்ட்பார்சர் மற்றும் எம்எஸ்டிபார்சர் போன்ற சார்பு பகுப்பாய்வு

கருவிகள் 25, 000 சொற்கள் கொண்ட அதே தரவுத்தொகுதியில் பயிற்சி அளிக்கப்பட்டபோது, மால்ட்பார்சரின் மதிப்பீடுச் சோதனை துல்லியம் 90.19% ஆகவும் MSTParser MSTParse-இன் மதிப்பீட்டுச் சோதனை துல்லியம் 81.13% ஆகவும் இருந்தது. சொல் அடையாளப்படுத்தி மற்றும் கூறாய்வியின் விளைவு சார்புப் பகுப்பாய்வு ஒழுங்குமுறையின் ஒட்டுமொத்த துல்லியத்தை பாதிக்கின்றன.

அரியரத்னமும் அவரது பங்காளிகளும் (Ariyarathnam et al 2014) இலங்கை-தமிழ் வாக்கியங்களிலிருந்து பயனுள்ள வாக்கியத் தகவல்களை மீட்டெடுப்பதற்காக இயற்கை மொழி வாக்கியங்களுக்கு ஒரு பகுதி கட்டமைப்பை ஒதுக்க வடிவமைக்கப்பட்ட ஆழமற்ற பகுப்பாய்வியை உருவாக்க முயற்சித்துள்ளனர். இது ஒரு வாக்கியத்தில் ஒவ்வொரு வார்த்தையையும் தானாகவே பொருத்தமான சொல்வகைப்பாட்டு அடையாளத்துடன் புலக்குறிப்பு செய்யும் சொல்வகை அடையாளப்படுத்தி மற்றும் அடையாளப்படுத்தப்பட்ட பெரிய தரவுத்தொகுதியின் தேவையின்றி வாக்கியங்களைத் தொடர்புபடுத்தப்பட்ட சொல் குழுக்களாக பிரிக்கும் விதி அடிப்படையிலான கூறாய்வி இவற்றின் அதிகபட்ச என்ட்ரோபி அடிப்படையிலான கலவையைப் பயன்படுத்துகிறது. இதைச் செய்ய, அவர்கள் உள்ளீட்டைப் பயன்படுத்தி 20 சொல்வகைப்பாடு அடையாளங்களைக் கொண்ட சொல்வகைப்பாட்டு அடையாளத் தொகுதியை உருவாக்கி, ஏறக்குறைய 12500 சொற்களைக் கொண்ட ஒரு தரவுதொகுதியை மனித முயற்சியால் அடையாளப்படுத்தி கூறுகளைப் பிரித்தெடுக்க 390 கூறு அமைபொழுங்குகளை அடையாளம் கண்டனர். சொல்வகை அடையாளப்படுத்தியும் கூறாய்வியும் முறையே 81.72% மற்றும் 78.3% f-அளவுகளை உறுதிப்படுத்தின. ஒருங்கிணைந்த ஆழமில்லாப் பகுப்பாய்வி 66.6% f-அளவை அளிக்கிறது.

12.3.2.3.10. எச்சத்தொடர் எல்லையை அடையாளம்காணல்

எச்சத்தொடர் என்பது ஒரு கருத்துரையை வெளிப்படுத்தக்கூடிய குறைந்தபட்ச இலக்கண அலகு ஆகும். இது வெளிப்படையான அல்லது மறைமுகமான இயல்புடைய ஒரு வினைச்சொல் அல்லது வினை குழு (வினை மற்றும் அதன் துணை வினை) மற்றும் அதன் பங்கெடுப்பாளர்கள் ஆகியவற்றைக் கொண்ட தொடர்ச்சியான சொற்களின் குழு ஆகும் (Vijay Sundar Ram and Sobha 2008). இது மொழி இலக்கணங்களில் ஒரு எச்சத்தொடரை ஒரு முக்கிய அலகாகக் காண்கின்றது மற்றும் மொழியியல் ஆய்வுகளின் ஒரு பகுதியாக அவற்றை அடையாளம் கண்டு வகைப்படுத்த வேண்டியதன் அவசியத்தை வலியுறுத்துகிறது. ஒரு எளிய வாக்கியத்துடன் ஒப்பிடும்போது சிக்கலான வாக்கியங்களின் பகுப்பாய்வும் செயற்பாங்கும் மிகவும் சவாலான

செயல்பாடாகும். வாக்கியத்தின் சிக்கலானது அதிகரிக்கும்போது இயற்கை மொழி ஆய்வுப் பயன்பாடுகள் பெரும்பாலும் மோசமாக செயல்படுகின்றன. ஒரு சிக்கலான வாக்கியத்தின் எச்சத்தொடர்கள் வாக்கியத்தில் அவற்றின் தொடரியல் செயல்பாடு அடிப்படையில் சரியாக அடையாளம் காணப்பட்டு வகைப்படுத்தப்பட்டால் அவ்வாக்கியத்தை ஆய்வது சாத்தியமற்றது ஆகும். இயந்திர மொழிபெயர்ப்பு, இணைத் தரவுத்தொகுதி ஒழுங்குபடுத்தல் (சீரமைப்பு), தகவல் பிரித்தெடுத்தல், தொடரியல் பகுப்பாய்வு, தானியங்கிச் சுருக்கம் மற்றும் பேச்சுப் பயன்பாடுகள் போன்ற பல இயற்கை மொழி ஆய்வு ஒழுங்கமைப்புகளின் செயல்திறன் ஒரு வாக்கியத்தில் எச்சத்தொடர் எல்லைகளை அறிமுகப்படுத்துவதன் மூலம் மேம்படுகிறது. AUKBCRC தமிழில் எச்சத்தொடர் எல்லை அடையாளம் காணும் ஒரு முன்னோடியான செயல்பாட்டை மேற்கொண்டுள்ளது (Vijay Sundar Ram and Sobha 2008; Vijay Sundar Ram et al 2012). அமிர்தா பல்கலைக்கழகத்தினர் சார்பு பகுப்பாய்வைப் பயன்படுத்தி தமிழ் மொழிக்கான எச்சத்தொடர் எல்லையை அடையாளம் காண முயற்சித்தனர் (Divya et al (2011).

12.3.2.3.11. ஒளிவழி எழுத்து உணர்தல் (Optical Character Recongizaton)

ஒளி எழுத்துணரி (Optical Character Recognition - OCR) என்பது எழுதப்பட்ட அல்லது பதிக்கப்பட்ட எழுத்துக்களை இயந்திர மூலமாகக் கணினிக்கு புரியும்படி மாற்றும் ஒரு நுட்பம் ஆகும். முதலில் ஆவணங்களை வருட வேண்டும் (scanning). இதற்கு நுண்மையான உணர் திறன் வாய்ந்த வருடி (scanner) தேவைப்படும். பின்னர் இதை மென்பொருள் மூலம் கணினிக்கும் புரியும்படி ஆக்கலாம்.

அச்சிடப்பட்ட உரையை குறிக்கும் ஒரு படத்தைக் கொடுத்தால், ஒளிவழி எழுத்துணரி தொடர்புடைய உரையைத் தீர்மானிக்கிறது. தமிழிற்காக பல ஒளிவழி எழுத்துணரி ஒழுங்கமைப்புகள் உருவாக்கப்பட்டுள்ளன. பலவகையான ஒளிவழி எழுத்துணரிகள் முயற்சிக்கப்படன: அச்சிடப்பட உரைகளை உணர்தல், கையெழுத்து உரைகளை அல்லது பிரதிகளை உணர்தல், சுவடி உரைகளை உணர்தல், பழைய ஆணவங்களின் உரைகளை உணர்தல் என்பன சிலவாகும். எழுத்துணரி உருவாக்கம் பற்றி பல கட்டுரைகள் வெளிவந்துள்ளன. இடங்கருதி இவற்றை இக்கட்டுரையில் தரவில்லை.

வி. கிருஷ்ணமூர்த்தி தமிழிற்கான ஒளி எழுத்துணரியின் உருவாக்கத்திற்குப் பெயர் பெற்றவர். அமிர்தா பல்கலைக்கழகம் சராசரி தரமுள்ள தமிழிற்கான ஒழிவழி எழுத்துணரியை உருவாக்க முயற்சித்தது. இந்திய மொழி தொழில்நுட்ப பெருக்கம் மற்றும் மேம்பாட்டு மையம் (Indian Language Technology Proliferation and Deployment Centre (ILTPDC) பதிவிறக்கம்

செய்ய இயலும் e -அக்ஷாராயன்-தமிழ் என்ற ஒளி எழுத்துணரியை உருவாக்கியுள்ளது. e-அக்ஷாராயன்-தமிழ் என்பது ஸ்கேன் செய்யப்பட்ட அச்சிடப்பட்ட இந்திய மொழி ஆவணங்களை யூனிகோட் குறியாக்கத்தில் முழுமையாக திருத்தக்கூடிய உரை வடிவமாக மாற்றுவதற்கான டெஸ்க்டாப் மென்பொருளாகும். இது விண்டோஸ் 7, 8 மற்றும் 10இல் செயல்பாடு செய்கிறது (டி.டி.ஐ.எல் வலைத்தளம்).

இந்திய அறிவியல் நிறுவனம் (Indian Institue of Science), பெங்களூர் தமிழுக்கு ஒரு முழுமையான ஒளி எழுத்துணரியை உருவாக்கி உள்ளது (Aparna and Ramakrishnan). இந்த தமிழ் ஒளி எழுத்துணரி ஒழுங்கமைப்பு எழுத்துரு (font) சுதந்திரமான மற்றும் அளவு (size) சுதந்திரமான சூழ்நிலையில் செயல்படுகிறது. துல்லியமான சாய்வைக் (skew) கண்டறிதல் அணுகுமுறை பயன்படுத்தப்பட்டுள்ளது; இது உண்மையான மதிப்பைப் பற்றி ± 0.06 வரம்பிற்குள் சாய்வு கோணத்தைக் கண்டறிய உதவுகிறது. அளவீட்டு விளைவுகளைத் தவிர்ப்பதற்கும், எழுத்தில் (character) இருக்கும் சிதைவுகளைக் குறைப்பதற்கும் சாம்பல் நிலை படத்தில் சாய்வு சுழற்சி செய்யப்படுகிறது. வடிவியல் தருணங்கள் மற்றும் தனித்துவமான கொசைன் உருமாற்றங்கள் ஆகியவை பயன்படுத்தப்படும் அம்சங்கள் ஆகும்; மற்றும் பயன்படுத்தப்படும் வகைப்படுத்தி அருகிலுள்ள அண்மையம் ஆகும். ஒட்டுமொத்த அறிதல் துல்லியம் சுமார் 98% ஆகும்.

கூகிள் தமிழிற்காக ஒப்பீட்டளவில் திறமையான ஆன்லைன் ஒளி எழுத்துணரி முறையை உருவாக்கியுள்ளது. இது மிகவும் நன்றாக வேலை செய்கிறது.

12.3.2.4. பொருண்மையியல் பகுப்பாய்வு

பொருண்மையியல் பகுப்பாய்வு பொருண்மையை ஆய்தல் மற்றும் புரிதல் என்பதைக் குறிப்பிடும். இது பின்வரும் சிக்கலான செயல்பாடுகளை உள்ளடக்கியது: ஒருபொருள் பன்மொழிகளைக் கண்டறிதல், சொல்மயக்க நீக்கம், கேள்வி-பதில் அமைப்புகளை உருவாக்குதல், ஒரு இயற்கை மொழியில் இருந்து மற்றொரு மொழிக்கு மொழிபெயர்ப்பது, அறிவின் அடிப்படையை பெருக்குதல். எந்தவொரு பொருண்மையியல் சிக்கலையும் தீர்க்க முயற்சிக்கும் முன், ஒருவர் உருபனியல் மற்றும் தொடரியல் பகுப்பாய்வை முடிக்க வேண்டும். வாசு அரங்கநாதனின் (2016ப்) கணினிசார் பொருண்மையியல் தமிழுக்கான குறிப்பிடத்தகுந்த பங்களிப்பாகும்.

12.3.2.4.1. சொற்பொருண்மை மயக்கநீக்கம்

சொற்களுக்கு பல்வேறு அர்த்தங்கள் உள்ளன. கொடுக்கப்பட்ட சூழலில் சரியான அர்த்தத்தைத் தேர்வுசெய்ய சொற்பொருண்மை மயக்கநீக்க ஒழுங்கமைப்பு நமக்கு உதவுகிறது.

இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்புகள், தகவல் மீட்டெடுப்பு ஒழுங்கமைப்புகள் மற்றும் கேள்விக்குப் பதிலளிக்கும் ஒழுங்கமைப்புகள் போன்ற பயன்பாடுகளுக்கு சொற்பொருண்மை மயக்கநீக்கம் மிகவும் முக்கியமானது. எடுத்துக்காட்டாக, ஆங்கிலத்திலிருந்து தமிழுக்கு ஒரு இயந்திர மொழிபெயர்ப்பு செய்கையில் 'bank' போன்ற பல்பொருள் ஒழுமொழிச் சொற்களின் சொற்பொருண்மை மயக்கத்தை நீக்கவேண்டும். அவ்வாறு செய்தால் தான் மூலமொழியை இலக்கு மொழியில் சரியாக மொழிபெயர்க்க இயலும் (நிகரான சரியான தமிழ்ச் சொல்லைத் தேர்ந்தெடுக்க இயலும்).

தமிழிற்கான சொற்பொருண்மை மயக்கநீக்க ஒழுங்கமைப்புகளை உருவாக்க பலர் முயற்சித்துள்ளனர். பாராட்டத்தக்க சொற்பொருண்மை மயக்கநீக்க ஒழுங்கமைப்பு உருவாக்கம் AUKBCRCஇல் செயல்படுத்தப்பட்டது. இது தமிழுக்கான ஒரு முழுமையான சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறையை உருவாக்கும் முயற்சியாக அமைந்தது (Baskaran, 2001). அமிர்தாப் பல்கலைகழகத்திலும் சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறையைகளை உருவாக்கும் முயற்சிகள் வெவேறு அணுகுமுறைகளில் மேற்கொள்ளப்பட்டன: 1. Tamil Word Sense Disambiguation using Support Vector Machines with rich features (Anand Kumar et al 2014a), 2. Supervised Cross-lingual Preposition Disambiguation for Machine Translation (Anand Kumar et al 2014b), 3. Computational approach to Word Sense disambiguation in Tamil (Rajendran et al 2013), 3. Corpus based approach for resolving verbal polysemy in Tamil (Rajendran et al 2014). சந்தோஷ்குமார் (2016) தமிழிலிருந்து ஆங்கில புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்புக்குப் பொருண்மை வலையைப் பயன்படுத்தி சொற்பொருண்மை மயக்கநீக்கம் செய்வதை முன்மொழிந்தார். பொருண்மை வலைத் தொழில்நுட்பத்தைப் பயன்படுத்துவதன் மூலம் தற்போதுள்ள புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு முறைகளை மேம்படுத்துவதற்கான ஒரு நுட்பத்தை அவர் முன்மொழிந்தார்.

12.3.2.4.2. கேள்விப்பதில் ஒழுங்கமைப்பு (Question-answer system)

கேள்விப்பதில் ஒழுங்கமைப்பு என்பது தகவல் மீட்டெடுப்பு மற்றும் இயற்கை மொழி ஆய்வு இவற்றுடன் தொடர்புடைய ஒரு கணினி மொழியியல் சார் முயற்சியாகும். இது இயற்கையான மொழியில் மனிதர்கள் எழுப்பும் கேள்விகளுக்குத் தானாகவே பதிலளிக்கும் கட்ட அமைப்புகளில் அக்கறை கொண்டுள்ளது. கேள்விப்பதில் நெறிமுறை, பொதுவாக ஒரு கணினி நிரல், அறிவு அல்லது தகவலின் கட்டமைக்கப்பட்ட தரவுத்தளத்தை வினவுவதன் மூலம் அதன் பதில்களை உருவாக்கலாம். கேள்விப்பதில் அமைப்புகள் இயற்கையான மொழி ஆவணங்களின்

கட்டமைக்கப்படாத தொகுப்பிலிருந்தும் பதில்களை எடுக்கலாம். கேள்விப்பதில் அமைப்புகளுக்குப் பயன்படுத்தப்படும் இயற்கை மொழி ஆவணச் சேகரிப்புகளின் சில எடுத்துக்காட்டுகள் பின்வருமாறு:

- குறிப்பு நூல்களின் உள்ளூர் தொகுப்பு
- உள் அமைப்பு ஆவணங்கள் மற்றும் வலைப்பக்கங்கள்
- தொகுக்கப்பட்ட நியூஸ்வைர் அறிக்கைகள்
- விக்கிபீடியா பக்கங்களின் தொகுப்பு
- உலகளாவிய வலைப்பக்கங்களின் துணைத்தொகுதி

கேள்வி பதில் ஒழுங்குமுறை பல வகையான கேள்வி வகைகளைச் சமாளிக்க முயற்சிக்கிறது: உண்மை, பட்டியல், வரையறை விளக்கம், 'எப்படி', ஏன்', கற்பனைகள், பொருண்மையியல் கட்டுப்பாட்டிற்கு உட்பட்ட செய்தி மற்றும் குறுக்கு-மொழி (மொழிகடந்த) கேள்விகள்.

இந்த ஒழுங்கமைப்பு ஒரு மனித மொழி கேள்விக்கு அதன் பதிலைத் தீர்மானிக்கும். வழக்கமான கேள்விகளுக்கு ஒரு குறிப்பிட்ட சரியான பதில் இருக்கும் ("இந்தியாவின் தலைநகரம் என்றால் என்ன?" போன்றவை); ஆனால் சில நேரங்களில் திறந்தநிலைக் கேள்விகளும் கேட்கப்படுகின்றன ("வாழ்க்கையின் பொருள் என்ன?" போன்றவை). சமீபத்திய கேள்விப்பதில் ஒழுங்கமைப்புகள் இன்னும் சிக்கலான கேள்விகளை நேரிடுகின்றன.

தமிழிற்காக ஒரு சில கேள்வி பதில் ஒழுங்கமைப்புகள் உருவாக்கப்பட்டுள்ளன. இவற்றில் குறிப்பிடத்தகுந்தவை அண்ணாப் பல்கலைக்கழகம், சென்னையிலிருந்து உருவாக்கப்பட "உருபனியல் பண்புக்கூறுகளைப் பயன்படுத்தி தமிழ் கேள்வி வகைப்பாடு" (Lakshmana Pandiyan & Geetha), விகெனேஷ் மற்றும் சௌவிமியாவின் "தமிழிற்கான தானியங்கி கேள்வி உருவாக்கி" (.Vignesh & Sowmya 2013) என்பனவாகும். மேலும் பலர் இத்தகைய கேள்விபதில் ஒழுங்கு முறைகளை உருவாக்கி இருக்கலாம். காலம் மற்றும் இடம் காரணமாக அவை இங்கு விடுபட்டுள்ளன.

12.3.2.4.3. உறவு பிரித்தெடுத்தல்

உறவு பிரித்தெடுத்தல் என்பது ஒரு உரையிலிருந்து பொருண்மை உறவுகளைப் பிரித்தெடுக்கும் செயல்பாடாகும். பிரித்தெடுக்கப்பட்ட உறவுகள் வழக்கமாக ஒரு குறிப்பிட்ட வகையின் இரண்டு அல்லது அதற்கு மேற்பட்ட இருப்புப்பொருள்களுக்கு (entities) இடையில் நிகழ்கின்றன (எ.கா. நபர், அமைப்பு, இருப்பிடம்) மற்றும் பல பொருண்மை வகைகளில் (எ.கா. திருமணம், வேலை, வேலை, வாழ்கின்றன) வருகின்றன. உரையின் ஒரு பகுதியைக் கொடுத்தால்,

உறவு பிரித்தெடுத்தல் பெயரிடப்பட்ட இருப்புப்பொருள்களுக்கு இடையிலான் உறவுகளை அடையாளம் காட்டுகிறது (எ.கா. யார் யாரை மணந்தார்). AUKBCRC இதற்கான முயற்சியில் ஈடுபட்டிருந்தது (Menaka et al 2014).

12.3.2.4.4. உணர்வுப் பகுப்பாய்வு (Sentiment analysis)

உணர்வுப் பகுப்பாய்வு என்பது ஒரு உரையில் வெளிப்படுத்தப்பட்ட கருத்துக்களை கணக்கீட்டு ரீதியாக அடையாளம் கண்டு வகைப்படுத்துவதற்கான செயல்முறையாகும்; குறிப்பாக ஒரு குறிப்பிட்ட தலைப்பு, படைப்பு போன்றவற்றில் எழுத்தாளரின் அணுகுமுறை நேர்மறையா எதிர்மறையா நடுநிலையா என்பதை தீர்மானிக்க உணர்வுப் பகுப்பாய்வு உதவும். உணர்வுப் பகுப்பாய்வு என்பது பொதுவாக ஆவணங்களின் தொகுப்பிலிருந்து அகநிலைத் தகவல்களைப் பிரித்தெடுப்பது ஆகும். பெரும்பாலும் நிகழ்நிலை (online) மதிப்புரைகளைப் பயன்படுத்தி குறிப்பிட்ட பொருள்களைப் பற்றிய "துருவமுனைப்பை" (polarity) தீர்மானிக்கிறது. சந்தைப்படுத்தல் நோக்கத்திற்காக, சமூக ஊடகங்களில் பொதுக் கருத்தின் போக்குகளை அடையாளம் காண இது மிகவும் பயனுள்ளதாக இருக்கும்.

உணர்வுப் பகுப்பாய்வு தமிழில் பரவலாக மேற்கொள்ளப்படுகிறது. வெவ்வேறு இலக்குகளை இலக்காகக் கொண்ட உணர்வுப் பகுப்பாய்வு மேற்கொள்ளப்படுகிறது. தமிழில் உணர்வுப் பகுப்பாய்வு அடிப்படையில் பல்வேறு வகையான முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. அமிர்த்தா விஷ்வவித்யபீடம் இதற்கான முயற்சிகளை மேற்கொண்டது. அருணும் பங்காளிகளும் (Arun et al 2015) ‘பண்புக்கூறு நிகழ்வெண் எண்ணிக்கை வழியாக தமிழ் திரைப்பட மதிப்புரைகளின் உணர்வுப் பகுப்பாய்வு’ என்ற உணர்வுப் பகுப்பாய்பாய்வை மேற்கொண்டனர். பத்மமாலாவும் பங்காளிகளும் (Padmamala 2017) ‘தமிழ் மொழிக்கான சுழல்நிலை நரம்பியல்பிணைய மாதிரிகள் அணுகுமுறையைப் பயன்படுத்தி ஆன்லைன் தமிழ் உள்ளடக்கங்களின் உணர்வுப் பகுப்பாய்வு’ என்ற உணர்வுப் பகுப்பாய்பாய்வை மேற்கொண்டனர். கௌசிகாவும் உமாவும் (2016) ‘ஆங்கிலம் மற்றும் தமிழ் ட்வீட்களின் உணர்வுப் பகுப்பாய்வு’ என்ற உணர்வுப் பகுப்பாய்பாய்வை மேற்கொண்டனர். இவை தவிர மேலும் பல உணர்வுப் பகுப்பாய்பாய்வுகள் மேற்கொள்ளப்பட்டன. காலம் இடம் கருதி அவற்றைப் பற்றிய விவரங்கள் இங்கு தவிர்க்கப்படுகின்றது.

12.3.2.4.5. தானியங்கி உரைச் சுருக்கம் (Automatic Text Summarization)

மூல ஆவணத்தின் முக்கிய விஷங்களை மட்டும் கொண்ட ஒரு சுருக்கத்தை உருவாக்க, மென்பொருள் மூலம் உரை ஆவணத்தை சுருக்கும் செயல்முறையே தானியங்கி சுருக்கம் ஆகும்.

ஒரு ஒத்திசைவான சுருக்கத்தை உருவாக்கக்கூடிய தொழில்நுட்பங்கள் நீளம், எழுதும் நடை மற்றும் தொடரியல் போன்ற மாறிகளைக் கணக்கில் எடுத்துக்கொள்கின்றன. தானியங்கித் தரவுச் சுருக்கம் இயந்திர கற்றல் மற்றும் தரவு அகழ்வாய்வு (data mining) இவற்றின் ஒரு பகுதியாகும். சுருக்கத்தின் முக்கியக் கருத்து முழுத் தொகுப்பின் "தகவலை" கொண்ட தரவின் துணைத் தொகுதியைக் கண்டுபிடிப்பதாகும். இத்தகைய நுட்பங்கள் இன்று தொழில்துறையில் பரவலாகப் பயன்படுத்தப்படுகின்றன. தேடுபொறிகள் ஒரு எடுத்துக்காட்டாகும். சுருக்கச் செயல்பாட்டில் ஆவணங்கள், படச் சேகரிப்புகள் மற்றும் வீடியோக்களின் சுருக்கமும் அடங்கும். ஆவணச் சுருக்கமானது முழுத் தகவல்தொடர்பு வாக்கியங்களைக் கண்டுபிடிப்பதன் மூலம் முழு ஆவணத்தின் பிரதிநிதித்துவ சுருக்கத்தை அல்லது சுருக்கத்தை உருவாக்க முயற்சிக்கிறது; அதே நேரத்தில் படச் சுருக்கத்தில் கணினி மிகவும் பிரதிநிதித்துவமான மற்றும் முக்கியமான (அதாவது முக்கிய) படங்களை கண்டுபிடிக்கும். கண்காணிப்பு வீடியோக்கள் மூலம் ஒருவர் நிகழ்வில்லாத சூழலிலிருந்து முக்கியமான நிகழ்வுகளைப் பிரித்தெடுக்க விரும்பலாம்.

தானியங்கி சுருக்கத்திற்கு இரண்டு பொதுவான அணுகுமுறைகள் உள்ளன: பிரித்தெடுப்பு (extraction) மற்றும் சுருக்கமாக்கம் (abstraction). சுருக்கத்தை உருவாக்குவதற்கு அசல் உரையில் இருக்கும் சொற்கள், சொற்றொடர்கள் அல்லது வாக்கியங்களின் துணைக்குழுவைத் தேர்ந்தெடுப்பதன் மூலம் பிரித்தெடுப்பு முறைகள் செயல்படுகின்றன. இதற்கு நேர்மாறாக, சுருக்கமாக முறைகள் அகம்சார்ந்த பொருண்மைசார் பிரதிநிதித்துவத்தை உருவாக்கும்; பின்னர் இயற்கையான மொழி உருவாக்கும் நுட்பங்களைப் பயன்படுத்தி ஒரு மனிதன் வெளிப்படுத்தக்கூடியவற்றுடன் நெருக்கமாக இருக்கும் சுருக்கத்தை உருவாக்கும். அத்தகைய சுருக்கத்தில் வாய்மொழிக் கண்டுபிடிப்புகள் இருக்கலாம். இன்றுவரை ஆராய்ச்சி முதன்மையாகப் பிரித்தெடுப்பு முறைகளில் கவனம் செலுத்தியுள்ளது. அது படச் சேகரிப்புச் சுருக்கம் மற்றும் வீடியோ சுருக்கத்திற்கும் பொருந்தும்.

தானியங்கு உரைச் சுருக்கமயமாக்கல் வழிமுறை உரையின் ஒரு பகுதியின் படிக்கக்கூடிய சுருக்கத்தை உருவாக்குகிறது. ஒரு செய்தித்தாளின் நிதிப் பிரிவில் உள்ள கட்டுரைகள் போன்ற அறியப்பட்ட வகை உரையின் சுருக்கங்களை வழங்க இது பெரும்பாலும் பயன்படுத்தப்படுகிறது.

பல்வேறு முறைகளைப் பயன்படுத்தி வெவ்வேறு வகையான தமிழ் உரைகளைத் தானாகச் சுருக்கமாக உருவாக்க ஒரு சில ஒழுங்கமைப்புகள் உருவாக்கப்பட்டுள்ளன. சென்னையிலுள்ள அண்ணாப் பல்கலைக்கழகம் (Banu et al 2007, AUKBCRC (Pattabi and Sobha 2017)., அமிர்தா விஷ்வவித்யபீடம் போன்ற நிறுவங்கள் இச்செயல்முறையில் ஈடுபட்டுள்ளன .

12.3.2.4.6. ஒத்ததைக் குறிப்பிடும் தீர்மானம் (Co-reference resolution)

ஒத்ததைக் குறிப்பிடும் தீர்மானம் என்பது ஒரு உரையில் ஒரே அமைப்பைக் குறிக்கும் அனைத்து வெளிப்பாடுகளையும் கண்டுபிடிக்கும் செயல்பாடாகும். ஆவணச் சுருக்கம், கேள்வி பதில் மற்றும் தகவல் பிரித்தெடுத்தல் போன்ற இயற்கையான மொழி புரிதலை உள்ளடக்கிய பல உயர் நிலை இயற்கைமொழி ஆய்வுச் செயற்பாடுகளுக்கு இது ஒரு முக்கியமான படியாகும்.

கொடுக்கப்பட்ட வாக்கியத்திலிருந்து அல்லது பெரிய உரையின் எந்தெந்த சொற்கள் (இருப்புப்பொருள்கள்) ஒரே பொருள்களைக் குறிக்கின்றன என்பதை ஒத்ததைக் குறிப்பிடும் அல்காரிதம் தீர்மானிக்கிறது. முற்சுட்டுத் தீர்மானம் (anaphora resolution) இந்தச் செயல்பாட்டு நடைமுறையின் ஒரு குறிப்பிட்ட எடுத்துக்காட்டாகும். மேலும் அவை குறிப்பிடும் பெயர்ச்சொற்கள் அல்லது பெயர்களுடன் பதிலீடுபெயர்களைப் பொருத்துவதில் குறிப்பாக அக்கறை கொண்டுள்ளது. ஒத்ததைக் குறிப்பிடும் தீர்மானத்தின் மிகவும் பொதுவான பணி, வெளிப்பாடுகளைக் குறிக்கும் "உறவுகளை இணைத்தல்" என்று அழைக்கப்படுவதையும் அடையாளம் காண்பது.

AUKBCRC இந்தச் செயல்பாட்டில் முன்னணி வகிக்கின்றது. சோபா ஒத்ததைக் குறிப்பிடும் தீர்மன ஒழுங்கமைப்பு உருவாகத்தின் முன்னணி ஆர்வலர் ஆவார். அவரது மேற்பார்வையின் கீழ் AUKBCRCஇல் இதுகுறித்த பல செயல்பாடுகள் மேற்கொள்ளப்பட்டுள்ளன. முற்சுட்டுத் தீர்மானம், மாற்றுப்பெயர் தீர்மானம் (pronominal resolution) போன்ற பரிமாணங்களில் ஆய்வுகள் மேற்கொள்ளப்பட்டன (Sobha et al 2014a, Sobha et al 2014b; Ahilandeswari and Sobha 2013, 2014 and 2015; Viajay Sanaka Ram and Sobha 2012, 2013 and 2016). மொழிகளைக் கடந்து ஒத்ததைக் குறிப்பிடும் ஆவணங்களை அடையாளம் காணவும் முயற்சி மேற்கொள்ளப்பட்டது (Pttabi and Sobha (2008).

12.3.2.3.7. பெயரிடப்பட்ட இருப்புப்பொருள் கண்டுபிடிப்பு (Named Entity Recognition (NER))

பெயரிடப்பட்ட இருப்புப்பொருள் கண்டுபிடிப்பு (Named Entity Recognition (NER)) என்பது சில குறிப்பிட்ட பண்புக்கூறுகள் கொண்ட இருப்புப்பொருள்களை அடையாளம் காண்பதாகும். இது இருப்புபொருள் கூறாய்வு மற்றும் இருப்புப்பொருள் பிரித்தெடுத்தல் என்றும் அழைக்கப்படுகிறது. இது தகவல் பிரித்தெடுத்தலின் (Information Extraction) ஒரு துணைச் செயல்படாகும். இது கட்டமைக்கப்படாத உரையில் உள்ள பெயரிடப்பட்ட இருப்புபொருள் குறிப்பீடுகளை நபரின் பெயர்கள், நிறுவனங்கள் இருப்பிடங்கள், மருத்துவக் குறியீடுகள், நேர

வெளிப்பாடுகள், அளவுகள், பண மதிப்புகள், சதவீதங்கள் போன்ற முன் வரையறுக்கப்பட்ட வகைகளாகக் கண்டறிந்து வகைப்படுத்த முயல்கிறது.

முழு பெயரிடப்பட்ட இருப்புபொருள் கண்டுபிடிப்பு கருத்தியல் ரீதியாகவும் செயல்படுத்தல்களிலும் பெரும்பாலும் இரண்டு தனித்துவமான சிக்கல்களாகக் காணப்படுகின்றது: பெயர்களைக் கண்டறிதல் மற்றும் அவை குறிப்பிடும் இருப்புப்பொருள்களின் வகை மூலம் பெயர்களை வகைப்பாடு செயதல் (எ.கா. நபர், அமைப்பு, இருப்பிடம் மற்றும் பிற). பெயரிப்பட்ட இருப்பொருள் கண்டுபிடிப்பு ஒழுங்குமுறை, ஒரு உரையில் இருக்கும் பெயரிடப்பட்ட இருப்புப்பொருள்களை "தனிநபர்கள்", "நிறுவனங்கள்", "இடங்கள்", "அமைப்பு", "நகரங்கள்", "தேதிகள்", "தயாரிப்பு" போன்ற முன் வரையறுக்கப்பட்ட வகைகளாக வகைப்படுத்துகிறது. இது ஒருவரின் உள்ளடக்க அறிவில் குறிபிட்ட பொருண்மையியல் அறிவைச் சேர்க்கிறது மற்றும் எந்தவொரு உரையின் விஷயத்தையும் உடனடியாகப் புரிந்துகொள்ள உதவுகிறது. பெயரிடப்பட்ட இருப்புப்பொருள் கண்டுபிடித்தல் மூலம் முழு உரையையும் தானாகவே ஒளிவருடல் செய்து, அவற்றில் விவாதிக்கப்பட்ட முக்கிய நபர்கள், நிறுவனங்கள் மற்றும் இடங்கள் எது என்பதை வெளிப்படுத்தலாம். ஒவ்வொரு உரைக்கும் பொருத்தமான அடையாளங்களை அறிந்துகொள்வது வரையறுக்கப்பட்ட படிநிலைகளில் உள்ள உரைகளைத் தானாக வகைப்படுத்த உதவுகிறது மற்றும் எளிதான உள்ளடக்கக் கண்டுபிடிப்பைச் செயல்படுத்த உதவுகிறது.

மொழியியல் இலக்கண அடிப்படையிலான நுட்பங்களையும் இயந்திர கற்றல் போன்ற புள்ளியியல்சார் மாதிரிகளையும் பயன்படுத்தும் பெயரிப்பட்ட இருப்பொருள் கண்டுபிடிக்கும் ஒழுங்கமைப்புகள் உருவாக்கப்பட்டுள்ளன. கையால் வடிவமைக்கப்பட்ட இலக்கண அடிப்படையிலான ஒழுங்கமைப்புகள் பொதுவாக சிறந்த துல்லியத்தைப் பெறுகின்றன; ஆனால் அனுபவம் வாய்ந்த கணினி மொழியியலாளர்களின் காலவிரையமும் குறைந்த மீட்பும் இதன் குறைபாடாகும். புள்ளியியல்சார் இருப்புப்பொருள் கண்டுபிடிக்கும் அமைப்புகளுக்குப் பொதுவாக கையால் அடையாளப்படுத்தப்பட்ட பயிற்சித் தரவு தேவைப்படுகிறது. கையால் அடையாளப்படுத்தும் முயற்சியைத் தவிர்க்கப் பகுதிமேற்பார்வை அணுகுமுறைகள் (semisupervised approaches) பரிந்துரைக்கப்பட்டுகின்றன. எந்திரக் கற்றல் அணுகுமுறை அடிப்படையிலான இருப்பொருள் கண்டுபிடிக்கும் ஒழுங்கமைப்பை உருவாக்கப் பல் வேறுபட்ட வகைப்படுத்தி வகைகள் (classifier types) பயன்படுத்தப்படுகின்றன.

இயற்கை மொழி ஆய்வு மற்றும் இயந்திர கற்றல் உள்ளிட்ட செயற்கை நுண்ணறிவின் (Artificial Intelligence (AI)) பல துறைகளில் பெயரிடப்பட இருப்பொருள் கண்டுபிடிப்பு பயன்படுத்தப்படுகிறது. உரை வரைபடத்தில் எவை மக்கள் அல்லது இடங்கள் போன்ற இயற்பெயர்களுக்குப் பொருந்தும் என்பதையும் மேலும் தரப்பட்ட உரையின் ஒழுக்கிலிருந்து அத்தகைய ஒவ்வொரு பெயரின் வகை என்ன (எ.கா. நபர், இருப்பிடம், அமைப்பு) என்பதையும் இருப்பொருள் கண்டுபிடிப்பு ஒழுங்கமைப்பு தீர்மானிக்கிறது.

தமிழ் மொழிக்கான பொதுவான இருப்பொருள் கண்டுபிடிப்பு ஒழுங்கமைப்புகளை உருவாக்குவதற்கு சில முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. இந்த முயற்சிகள் இயந்திர கற்றல் அணுகுமுறைகளான ஹிட்டன் மார்க்கோவ் மாடல் (Hidden Markov Model (HMM)), மாக்சிமம் என்ட்ரோபி மார்கோவ் மாடல் (Maximun Entoropy Makov Models), சப்போர்ட் வெக்டர் மெஷின் (Support Vector Machine) மற்றும் கண்டிஷனல் ராண்டம் ஃபீல்ட் (Conditional Radom Fields (CRF)) போன்றவற்றைப் பயன்படுத்துகின்றது.

தமிழில் பெயரிடப்பட்ட இருப்பொருள் கண்டுபிடிப்பு ஒழுங்கமைப்பின் உருவாக்கம் குறித்த ஒரு முன்னோடியான செயல்பாட்டை ஏ.யூ.கே.பி.சி.ஆர்.சி மேற்கொண்டது (Vijaykrishana and Sobha 2008, Malarkodi et al 2016; Malarkodi and Sobha 2018; Vijaykrishna and Sobha 2008). அமிர்தா பல்கலைக்கழகமும் தமிழிற்கான இருப்பொருள் கண்டுபிடிப்பு ஒழுங்கமைப்பை உருவாக்கப் பங்களிப்பு செய்துள்ளது (Abinaya et al 2014).

12.3.2.3.8.பிற கணினிப் பொருண்மையியல் ஆய்வுகள்

கணினிப்பொருண்மையியல் மிக விரிவான புலமாகும். கணினிப் பொருண்மையியல் அடிப்படையில் சில ஆய்வுகள் மேற்கொள்ளப்பட்டன. கணினிப்பொருண்மையியல் ஏ.யூ,கே.பி.சி.ஆர்.சி. மேற்கொண்கொண்ட ஆய்வுகள் இங்குக் குறிப்பிடத்தக்கன. இதில் மொழி ஆய்வுக்கான பொருண்மையியல்சார் அடையாளப்படுத்தல் (Arulmozhi, Sobha. 2006). மற்றொன்று காரண உறவுகள் மற்றும் அதன் தானியங்கி அடையாளம் காணல் (Menaka et al (2014).

12.3.2.4. உரை உருவாக்கம்

கணினி தரவுத்தளங்கள் அல்லது பொருண்மை நோக்கங்களிலிருந்து தகவல்களைப் படிக்கக்கூடிய மனித மொழியாக மாற்றுவது உரை உருவாக்கம் என்று அழைக்கப்படுகிறது. உரையை உருவாக்கும் ஒழுங்கமைப்பு அல்லது வழிமுறை உரை உருவாக்கி (text generatore) என்று அழைக்கப்படுகிறது. ஒரு பொதுவான உரை உருவாக்க ஒழுங்கமைப்பின் நோக்கம்,

முன்னரே கூறப்பட்ட சில குறிக்கோள்களைத் திருப்திப்படுத்தும் உரையை உருவாக்குவதாகும். இத்தகைய அமைப்புகள் ஒரு அறிவுத் தளம் மற்றும் குறிக்கோள்களின் தொகுப்பு இவற்றுடன் வழங்கப்படுகின்றன. அறிவுத் தளத்தில் வெளிப்படுத்தப்பட வேண்டிய தகவல்கள் உள்ளன. கணினி பின்னர் இந்தத் தகவலை வாக்கிய நீளக் கூறுகளாக ஒழுங்கமைக்கிறது, இந்தக் கூறுகளை வாக்கியங்களாக உணர்ந்து உரையை அச்சிடுகிறது அல்லது பேசுகிறது.

உரை உருவாக்கும் ஒழுங்கமைப்புச் செயல்பாட்டில் AUKBCRCயின் முயற்சி குறிப்பிடத்தகுந்தது. ஆனந் ராமகிருஷ்ணன் "மெல்லிசைகளுக்கான தமிழ் பாடல்களின் தானியங்கி உருவாக்கம்" என்பதை முயன்றார் (Anand Ramakrishnan et al 2009; Anand Ramakrishnan and Sobha 2010). சென்னை அண்ணாப் பல்கலைக்கழகமும் உரை உருவாக்கத்தில் ஈடுபட்டது. கோஹிலவாணியும் பங்காளிகளும் (Kohialavani et al 2009) ஒரு தானியங்கி தமிழ் உள்ளடக்க உருவாக்க ஒழுங்கமைப்பை அமைக்க முயன்றனர். தானியங்கி உள்ளடக்க உருவாக்க அமைப்பு தமிழ் மொழியில் அறிவார்ந்த பயிற்சி முறையை உருவாக்குவதை நோக்கமாகக் கொண்டுள்ளது. இந்த அமைப்பு தமிழ் மொழியில் தனிப்பயனாக்கப்பட்ட உள்ளடக்கத்தை ஒரு தனிப்பட்ட பயனர் தேவைகளுக்கு அவர்களின் கற்றல் திறன்கள் மற்றும் ஆர்வங்களின் அடிப்படையில் வழங்குவதில் கவனம் செலுத்துகிறது. ராஜேஸ்வரியும் பங்காளிகளும் (2017) 'குறிப்புகளிலிருந்து தமிழ் கதைகளைத் தானாக உருவாக்குவதற்கான மொழி உறவு மாதிரியை' உருவாக்குவதைப் பற்றி கூறுகின்றனர்.

12.3.2.5. இயந்திர மொழிபெயர்ப்பு

இயந்திர மொழிபெயர்ப்பு என்பது உரையை அல்லது பேச்சை மொழிபெயர்க்க மென்பொருளைப் பயன்படுத்துவதை ஆராயும் கணினி மொழியியலின் துணைத் துறையாகும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு மனித மொழியிலிருந்து மற்றொரு மொழிக்கு உரையைத் தானாக மொழிபெயர்க்கும். இயந்திர மொழிபெயர்ப்பு, சில நேரங்களில் எம்டி (Machine Translation (MT)என்ற சுருக்கத்தால் குறிப்பிடப்படுகிறது. இது கணினி உதவி மொழிபெயர்ப்பு, இயந்திர உதவியுடன் மனித மொழிபெயர்ப்பு, தானியங்கி இயந்திர மொழிபெயர்ப்பு, ஊடாடும் மொழிபெயர்ப்பு என வேறுபடும்.

ஒரு அடிப்படை மட்டத்தில், இயந்திர மொழிபெயர்ப்பு ஒரு மொழியின் சொற்களுக்கு மற்றொரு மொழியில் எளிமையான மாற்றீட்டைச் செய்கிறது, ஆனால் அது வழக்கமாக ஒரு உரையின் நல்ல மொழிபெயர்ப்பை உருவாக்க இயலாது; ஏனெனில் இலக்கு மொழியில் முழு சொற்றொடர்களின் புரிதலும் அவற்றின் இணைகளும் தேவை. தரவுத்தொகுதி (விரிதரவு)

புள்ளியியல் மற்றும் நரம்பியல் நுட்பங்களுடன் இந்தச் சிக்கலைத் தீர்ப்பது வேகமாக வளர்ந்து வரும் ஒரு துறையாகும். இது சிறந்த மொழிபெயர்ப்புகளுக்கு வழிவகுக்கிறது.

ஒரு மொழியின் அலகுகளுக்கு (வாக்கியம், எச்சத்தொடர்கள், தொடர்கள் சொற்கள், போன்ற அலகுகளுக்கு) நிகரான அலகுகளை மற்றொரு மொழியில் கண்டுபிடிப்பது மிகவும் கடினமான சிக்கல்களில் ஒன்றாகும். இதற்கு மனிதர்கள் வைத்திருக்கும் (இலக்கணம், சொற்பொருள், உண்மையான உலகத்தைப் பற்றிய உண்மைகள் போன்ற) பல்வேறு வகையான அறிவு தேவைப்படுகிறது.

12.3.2.6.1. தமிழுக்கு அனுசாரகா

அனுசாரகா சட்டகத்தின் செயல்முறையை அடிப்படையாகக் கொண்ட இந்தி-தமிழ் மொழிபெயர்ப்பிற்காக ஐ.ஐ.டி கான்பூரில் ராஜேந்திரன் அனுசாரகா என்ற மொழிபெயர்ப்புக் கருவியை உருவாக்கினார். பின்னர் இந்தச் செயல்பாடு AUKBCRC ஆல் மேற்கொள்ளப்பட்டு முடிக்கப்பட்டது.

12.3.2.6.2. அனுவாதக்–ஆங்கிலத்திலிருந்து இந்திய மொழிமொழிக்கு இயந்திர மொழிபெயர்ப்பு செய்யும் ஒரு வலுவான ஒழுங்குமுறை

இது கிளையமைப்பு இணைக்கும் இலக்கணம் (Tree Adjoing Grammar (TAG/டாக்) அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு, புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Statistical Machine Translation (SMT), விதிகள் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (AnalGen/அனல்ஜென்), மற்றும் எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Example Based Machine Translation (EBMT/ஈபிஎம்டி) ஆகியவ இது நான்கு இயந்திர மொழிபெயர்ப்பு தொழில்நுட்பங்களின் ஒருங்கிணைப்பை முன்வைத்துள்ள கூட்டமைப்பு நிறுவனங்களின் கூட்டு முயற்சி ஆகும். ஒரு நிபுணத்துவமான ஆங்கிலத்திலிருந்து இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்பு அமைப்பு (English to Indian Language Machine Translation (EILMT)) என்பது ஒரு அதிநவீன தீர்வாகும், இது உரையை ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு மொழிபெயர்க்க அனுமதிக்கிறது.

அனுவாதக்வ்ஷ் இயந்திர மொழிபெயர்ப்பு சி.டி.ஐ.சி./சிடாக் புனே தலைமையிலான 10 இந்திய கல்வி மற்றும் ஆராய்ச்சி நிறுவனங்களின் கூட்டமைப்பால் உருவாக்கப்பட்டது. இந்திய அரசின் மின்னணு மற்றும் தகவல் தொழில்நுட்பத் துறையின் (Departmet of Electronic and Communication (DeitY/டீடி) கீழ் 'இந்திய மொழிகளுக்கான தொழில்நுட்ப மேம்பாடு' (Technological Development of Indian Languges (TDIL/டி.டி.ஐ.எல்) என்ற திட்டத்தின்

ஆதரவுடன் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்புகள் 6 மொழி ஜோடிகளுக்கு அதாவது ஆங்கிலத்திலிருந்து இந்தி, பெங்காலி, மராத்தி, உருது, தமிழ் மற்றும் ஒரியா ஆகிய மொழிகளுக்கு ஒழுங்கமைப்புகள் உருவாக்கப்படுகிறன. ஆரம்பத்தில் சுற்றுலா மறும் உடல்நிலப் பொருண்மைக் களங்கள் எடுத்துக்கொள்ளப்பட்டன. பின்னர் படிப்படையாக பிற பொருண்மைக்களத்திற்கு விரிவுபடுத்தும் நோக்கம் கொண்டுள்ளது.

12.3.2.6.3. இந்திய மொழியிலிருந்து இந்திய மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பு (ILILMT Systems)

DeitYஆல் (Departmet of Electronic and Communication (DeitY/டீடி) நிதியளிக்கப்பட்ட கூட்டமைப்புத் திட்டம் இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு அமைப்புகளை உருவாக்கியுள்ளது. இந்த திட்டத்தின் கீழ் ஒரு சில தமிழ் சார்ந்த இயந்திர மொழிபெயர்ப்பு அமைப்புகள் உருவாக்கப்பட்டுள்ளன. அவை இந்தி-தமிழ், தமிழ்-இந்தி மொழிபெயர்ப்பு அமைப்பு, தெலுங்கு-தமிழ், தமிழ்-தெலுங்கு மொழிபெயர்ப்பு மற்றும் மலையாள-தமிழ், தமிழ்-மலையாளம் மற்றும் இயந்திர மொழிபெயர்ப்பு அமைப்பு என்பனவாகும். இந்த ஒழுங்கமைப்புகள் மேலும் மேம்படுத்தப்படவேண்டும். நீதியின்மை காரணமாக இம்முயற்சிகள் தற்காலிகமாகக் கைவிடப்பட்டுள்ளன.

12.3.2.6.4. இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் அமிர்தா விஷ்வ வித்யபீடத்தின் பங்களிப்புகள்

கோயம்புத்தூர் அமிர்தா பல்கலைக்கழகம் தமிழ் சார்ந்த எம்டி அமைப்புகளை உருவாக்குவதற்கான பல திட்டங்களை முன்வைத்துள்ளது. இது CDAC, புனேயின் தலைமையில் நடைபெற்ற EILMT கூட்டமைப்பு திட்டத்தின் கீழ் ஆங்கிலத்திலிருந்து தமிழுக்கு இயந்திர மொழிபெயர்ப்பை மேற்கொண்டு முடித்தளித்தது. இது TAG இலக்கணத்தை அடிப்படையாகக் கொண்டது.

ஆனந்த் குமார் (Anand Kumar 2015) தனது முனைவர் பட்ட ஆய்வின் ஒரு பகுதியாகக் கணினி மொழியியலின் அடிப்படையில் ஆங்கிலம்-தமிழ் கருத்துப் பரிமாற்றத்திற்கான புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பை (Stastical Machine Translation System) உருவாக்கினார். மேலும் ஆங்கிலத்திலிருந்து தமிழ் மொழிக்கு காரணியாலான புள்ளியியல் இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பை (Factored Statisticsl Machine Translaton System) முன்மொழிந்தார் (Anand Kumar et al 2010). மேலும் தமிழ்

இணையக் கல்விநிறுவனத்தின் நிதியுதவியுடன் "தமிழ் கற்பிப்பதற்கான கணினிக் கருவிகள்" (Computational Tools for Teaching Tamil) திட்டத்தின் கீழ், அமிர்தா பல்கலைக்கழகம் ஆங்கிலம்-தமிழ் இயந்திர மொழிபெயர்ப்பிற்கான விதி அடிப்படையிலான மொழிபெயர்ப்பு ஒழுங்கமை உருவாக்கியுள்ளது (Rajendran and Anand Kumar 2018). இதில் ஆங்கிலத்தின் ஸ்டான்போர்ட் பல்கலைக்கழக சார்பு பகுப்பாய்வியின் வெளியீடு தமிழ் சார்பு பகுப்பாய்வு வெளியீடாக மாற்றப்பட்டு மற்றும் உருபனியல்-தொடரியல் உருவாக்கியைப் பயன்படுத்தி தமிழ் உரை உருவாக்கப்படுகிறது. இந்த அமைப்பு அதன் தொடக்க நிலையில் உள்ளது. அது மேலும் மேம்படுத்தப்பட வேண்டும்.

12.3.2.6.5. இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் AUKBCRC இன் பங்களிப்பு

இந்தி-தமிழ் இயந்திரமொழிபெயர்ப்பு உதவிக்கருவி உருவாக்க ஐஐடி கான்பூரில் தொடங்கப்பட்ட அனுசரகா திட்டத்தை AUKBCRC நிறைவு செய்தது. இந்தியமொழியிலிருந்து இந்தியமொழி இயந்திரமொழிபெயர்ப்பு கூட்டமைப்பு திட்டத்தின் (consortium project) கீழ் இந்தி-தமிழ், தமிழ்-இந்தி இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பை உருவாகியது. இது தவிர, AUKBCRC ஆங்கிலம்-தமிழ் மொழிபெயர்ப்பிற்கான ஒரு இயந்திரமொழிபெயர்ப்பு ஒழுங்கமைப்பையும் உருவாக்கியுள்ளது. இது ஒரு விதி அடிப்படையிலான ஒழுங்கமைப்பு ஆகும்,

12.3.2.6.6. இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் தமிழ் பல்கலைக்கழகத்தின் பங்களிப்பு

தமிழ்ப் பல்கலைக்கழம் தான் முதன் முதலில் ஒரு ஒழுங்கான இயந்திர மொழிபெயர்பில் ஈடுபட்டது. இதற்கு முக்கிய காரணமாக இருந்தவர் அன்றைய துணைவேந்தர் பேராசிரியர் வி.ஐ. சுப்பிரமணியம் ஆவார். இவர் மேற்பார்வையில் ரஷ்ய-தமிழ் மொழிபெயர்ப்பு முயற்சி மேற்கொள்ளப்பட்டது (Chellamuthu et al 1984). மேலும் ஆங்கில மொழியியல் பாட நூல்களை தமிழில் மொழிபெயர்க்க ஒரு இயந்திர உதவிக்கருவியைத் தயாரிக்க முயற்சி எடுக்கப்பட்டது (Rajendran and Kamashi 2004).

12.3.2.6.7. உத்தமத்தின் பங்களிப்பு

தமிழ் தொடர்பான 'இயற்கை மொழி ஆய்வு மற்றும் கணிமொழியியல் ஆய்வின் வளர்ச்சியும் சவால்களும்' பற்றி கூறுகையில் உத்தமத்திம் பங்களிப்பு பற்றி கூறாமல் செல்லவியலாது. பின்வரும் தகவல்கள் அவர்களின் இணையதளத்திலிருந்து எடுக்கப்பட்டதாகும்:

உத்தமம் (International Forum for Information Technology in Tamil (INFITT) என்பது உலகளாவிய இலாப நோக்கற்ற அரசு சார்பற்ற நிருவனம். தமிழ் கம்ப்யூட்டிங்கை ஊக்குவிக்க அர்ப்பணிப்பணிக்கப்பட்ட அமைப்பு. அமெரிக்காவின் கலிபோர்னியாவில் அதிகாரப்பூர்வமாக

NPOஆக பதிவு செய்யப்பட்டுள்ளது. உத்தமத்தின் (INFITT) நோக்கம் “தமிழ் தகவல் தொழில்நுட்பம் மற்றும் அதனுடன் தொடர்புடைய தரங்களை உருவாக்க உதவுதல்”, “தமிழ் ஐடி டெவலப்பர் சமூகத்திற்காக தொழில்நுட்ப மன்றங்கள் மற்றும் வருடாந்திர சந்திப்புகள் /மாநாடுகள் வழங்குதல்.”, “பிராந்திய விருப்பமுள்ள குழுக்கள், தேசிய அரசு நிறுவனங்கள் இடையே ஒரு முக்கிய இணைப்பாக செயல்படுதல்”, ”எம்.என்.சி மற்றும் தரப்படுத்தல் நிறுவனங்களுக்கு தொழில்நுட்ப நிபுணத்துவ ஆதரவை வழங்குதல்”, “முன்னேற்றங்களுக்கு பொதுமக்களிடையே விழிப்புணர்வை ஏற்படுத்துதல்/ அதிகரித்தல், தமிழ் தகவல் தொழில்நுட்பப் பட்டறைகளை ஏற்பாடு செய்வதன் மூலம் இளைய திறமைகளை ஈர்த்தல்” என்பவாகும்.

உத்தமம் ஆண்டுதோறும் வெளியிடும் மாநாட்டுக் கட்டுரைகளின் தொகுப்புகள் ’இயற்கை மொழி ஆய்வு மற்றும் கணிமொழியியல் ஆய்வு’ அடிப்படையில் தமிழ் செய்யப்பட்டுவரும் ஆய்வுப்பணிகளையும் தமிழின் தொழில்நுட்ப வளர்ச்சியையும் பறைசாற்றும். பலருக்கும் தமிழின் தொழில்நுட்ப வளர்ச்சியில் தங்கள் பங்களிப்பைக் கூற இம்மாநாடுகளும் மாநாட்டுக் கட்டுரைகளின் தொகுப்புகளும் வகை செய்கின்றன.

12.5. மதிப்பீடு

இயற்கைமொழி ஆய்வு, கணினிமொழியியல் பின்னணியில் தமிழின் தொழில்நுட்டப முன்னேற்றதை அறிய வேண்டி முன்னர் கூறிய கணினிசார் ஒழுங்குமுறைகளின் எண்ணிக்கை, கிடைகும் நிலை, தரம், செயல் எல்லை, முழுமை/முதிர்ச்சி, ஈடுகொடுத்தல், மாற்றி அமைத்துக்கொள்ளும் திறன் என்ற அளகைகள் அடிப்படையில் ஒரு மதிப்பீட்டு அட்டவணை தயாரிக்கும் முயற்சி இங்கு மேற்கொள்ளப்பட்டுள்ளது. பின்வரும் அட்டவணை இந்திய மொழிகளின் தொழிநுட்ட முன்னேற்றத்தைக் குறித்து கட்டுரை எழுதுவதற்காக கிரிஷ்நாத் ஷா (Girish Nath Jah) “Instructions for writing white paper for Indian languages” என்ற தலைப்பில் அனுப்பிய அறிவுறுத்தல்களின் ஒரு பகுதியாகும். ஆனால் வேண்டிய தகவல்கள் இல்லாத காரணத்தால் இவ்வட்டவணையைப் பூர்த்தி செய்வது கடினம். இருப்பினும் ”இயற்கை மொழி ஆய்வு மற்றும் கணினிமொழியியலின் வளர்ச்சி” என்ற நிலையில் தமிழ் மொழியின் தொழில்நுட்டப வளர்ச்சியைப் பட்டியலிடும் போது இப்பட்டியலில் குறிப்பிடப்பட்டுள்ள கேள்விகள் எழும் என்ற காரணத்தால் முழுவதும் நிரப்படாத இந்தப் பட்டியல் கீழே தரப்பட்டுள்ளது.

	எண்ணிக்கை	கிடைக்கும்	தரம்	செயல் எல்லை	முழுமை/ முதிர்ச்சி	ஈடு கொடுத்த	மாற்றி அமைத்

		நிலை				ல்	துக்கொ ள்ளும் திறன்
மொழித் தொழில்நுட்பம்: கருவிகள், தொழில்நுட்பங்கள் மற்றும் பயன்பாடுகள்							
பேச்சுப்புரிந்த்து கொள்கை	6	4	-	-	-	-	-
பேச்சு உருவாகம்	7	6	-	-	-	-	-
இலக்கணப் பகுப்பாய்வு							
சொற்களாகப் பிரித்தல்	6	4					
பகுதி ஆய்வு	3	3					
சொல்லன் ஆய்வு	3	2					
உருபனியல் பகுப்பாய்வு	10	7					
உபனியல் உருவாக்கம்	6	3					
சொல்வகை பாடு அடையாளப்படு த்தல்	10	7					
தொடரியல் பகுப்பாய்வு	4	2					
கூறாய்வு	4	3					
ஆழமில்லாப் பகுப்பாய்வு	2	1					

எச்சத்தொடர் எல்லை அறிதல்	2	1					
ஒளி எழுத்துணரி	7	3					
பொருண்மையியல் ஆய்வி							
சொல் மயக்கநீக்கம்	4	2					
கேள்விப்பதில் ஒழுங்குமுறை	2	1					
உறவுகளைப் பிரித்தெடுத்தல்	3	2					
உணர்வுப் பகுப்பாய்வு	5	3					
தானியக்க உரை சுருக்கல்	4	3					
உரை உருவாகம்	2	1					
இயந்திர மொழிபெயர்ப்பு	10	5					
மொழியியல் வழங்கள்: தரவு மற்றும் அறிவு தளம்							
உரை தரவுதொகுப்பு	4	4					
பேச்சு தரவுத்தொகுப்பு	2	2					
இணைத் தரவுத்தொகுதி	4	3					
சொல்சார் தரவுத்தொகுதி	10	6					

இலக்கணங்கள்	6	6					

12.6. எதிர்கால ஆய்வு மற்றும் வளர்ச்சிக்கான செயல்பாடுகள்

தமிழ் இப்போது மத்திய அரசால் எதிர்பார்க்கின்றபடி முழு அளவிலான டிஜிட்டல் மயமாக்கலுக்குத் தயாராக உள்ளது. நிச்சயமாக இன்னும் முடிக்கப்படாத பல செயல்பாடுகள் உள்ளன. எடுத்துக்காட்டாக, இயந்திர மொழிபெயர்ப்பு அமைப்புகள் பொதுப் பயனர்களுக்கு ஏற்றவாறு மாற்றியமைக்கப்படவில்லை. தமிழ் சொல்வலை ஆங்கிலம் போன்ற ஐரோப்பிய மொழிகளுக்கு இணையாக விரிவாக்கப்பட வேண்டும் அல்லது அதிகரிக்கப்பட வேண்டும். தமிழுக்கு ஒதுக்கப்பட்ட யூனிகோட் இடங்களில் பல சிக்கல்கள் உள்ளன. யூனிகோட் முறையில் சரியான இலக்கண பகுப்பாய்விற்குத் தேவையான அளவுக்குத் தமிழருக்கு அதிக இடங்கள் தரப்படவில்லை. இன்னும் ஆய்வாளர்கள் மற்றும் பயனர்கள் தமிழை ரோமானாகவும் பின்னர் இலக்கண பகுப்பாய்விற்குப் பிறகு தமிழாகவும் மாற்றுகிறார்கள். தமிழை உள்ளிடுவதற்கு ஏற்றுக்கொள்ளப்பட்ட WX ஒலிபெயர்ப்பு முறை மிக மோசமானது. இந்தச் சிக்கல் தீர்க்கப்பட வேண்டும். இதுவரை உருவாக்கப்பட்ட பல மொழிசார் கருவிகள்/மென்பொருள்கள் திறந்த மூலமாக கிடைக்க வேண்டும். உரை தரவுத்தொகுதி, பேச்சுத் தரவுத்தொகுதி, இணைத் தரவுத்தொகுதி போன்ற மொழி வளங்களை பகிர்ந்து கொள்ளவும் பொது பயனர்களுக்கு கிடைக்கச்செய்யவும் வழிசெய்ய வேண்டும். முயற்சிகளைக் குறைக்க கிளவுட் சோர்சிங்கை ஊக்குவிக்க முடியும். தேவையற்ற 'செய்ததையே திருப்பிசெய்தல்' தவிர்க்கப்பட வேண்டும். ஒருவருக்கு கிடைக்கும் வளங்கள் பிறருடன் பகிர்ந்து கொள்ளப்பட வேண்டும். தமிழின் தொழில்நுட்ப வளர்ச்சிக்கு ஒருங்கிணைந்த முயற்சிகள் தேவை. கணினி அறிவியலாரும் மொழியியலாரும் மொழி வல்லுனர்களும் இணைந்து செயல்பட வேண்டும்..

ஐரோப்பிய மொழிகளுடன் ஒப்பிடும்போது தமிழின் தொழில்நுட்ப வளர்ச்சிக்குச் செய்யப்பட்டுள்ள செயல்பாடுகள் ஒப்பீட்டளவில் குறைவு. விரும்பிய இலக்கை அடைய நாம் நீண்ட தூரம் செல்ல வேண்டும். தனியார் துறை நிறுவனங்கள் உட்பட அனைத்து நிறுவனங்களும் இதற்கு உழைக்க வேண்டும். கணினி அறிவியலார்கள் மற்றும் மொழியியலாளர்கள் (மொழி வல்லுநர்கள் உட்பட) ஒரு ஒருங்கிணைந்த மற்றும் காலம் அடிப்படையில் திட்டமிடப்பட்ட இலக்குடன் கைகோர்த்து செயல்பட வேண்டும். இப்போது கூகிள் தமிழுக்கு இயற்கைமொழி ஆய்வுசார் ஒழுக்கமைப்புகளைக் கொண்டுள்ளது. எடுத்துக்காட்டாக மொழிபெயர்ப்பாளர்கள், ஓ.சி.ஆர்கள் போன்றவற்றைக் கூறலாம். ஆனால் கூகிள் புள்ளியியல் முறையால் செயல்படுகிறது, இது எப்போதும் நம்பகமானதாக இருக்காது. மேலும், கூகிளைச் சார்ந்திருப்பது எதிர்விளைவில்

முடியலாம். எனவே நமது மொழியின் தொழில்நுட்ப வளர்ச்சிக்கு நாமே மொழிக்கருவிகள், மென்பொருள்கள் வளங்களை உருவாக்க வேண்டும், தமிழ் மாணவர்களே நமது பலம். தமிழ் இயற்கைமொழி ஆய்விலும் கணினிமொழியியலிலும் பணியாற்ற அவர்களுக்கு பயிற்சி தந்து ஊக்குவிக்க வேண்டும். தமிழின் தொழில்நுட்ப வளர்ச்சிக்கு மத்திய, மாநில அரசுகள் நிதியளிக்க வேண்டும். வரி செலுத்துவோரின் பணத்தை நாம் மிகுந்த கவனத்துடன் பயன்படுத்த வேண்டும்.

கவனத்திற்கு: இக்கட்டுரை எனக்குக் கிடைத்த தகவல்கள் அடிப்படையில் எழுத்தப்படுள்ளது. பல பங்களிப்புகளும் கட்டுரைகளும் விடுபட்டிருக்கக் கூடும். குறிப்பாக வெளிநாட்டினரின் பங்களிப்புகள் முழுமையாகக் கூறப்படவில்லை. தமிழ் இணைய மாநாட்டுக் கட்டுரைகள் முழுவதுமாகப் பயன்படுத்தப்படவில்லை. அவைகுறித்து இக்கட்டுரையாளருக்கு மின் அஞ்சல் செய்தால் இக்கட்டுரையில் சேர்க்கவியலும்.

நோக்கீட்டு நூல்கள்

இராஜேந்திரன், ச. 1991.'தற்காலத் தமிழ்ச் சொற்களஞ்சியம் (பாகம் 1),' தமிழ்க்கலை, 9.1. -4:55-76.

இராசேந்திரன், ச. 1999. பொருட்புல வகைப்பாடும் சொற்களஞ்சியமும் [Classification into semantic domains and Thesaurus]. புலமை தொகுதி 25, எண்2, டிசம்பர், 1999, 47-66.

இராசேந்திரன், ச. 2001 தற்காலத் தமிழ்சொற்களஞ்சியம். தமிப்பல்கலைக் கழகம், தஞ்சாவூர்.

இராஜேந்திரன், ச. 2002. 'இணையம் வழி அயல்நாடுகளில் தமிழ் கற்றலை வளர்த்தல், 'தமிழ் ஆய்வில் இன்றைய போக்குகள், 75-81, திருமதி செண்பகம் சுப்பையா அறக்கட்டளை, தஞ்சாவூர்.

இராஜேந்திரன், ச. 2002.மொழி ஆய்வில் கணிப்பொறியின் பயன்பாடு [use of computer in language analysis]. சித்திரபுத்திரன் மற்றும் பிறர் (பதிப்பு). ஆய்வியல் கட்டுரைகள். அன்னைஅகரம் பதிப்பகம். தஞ்சாவூர்.

இராஜேந்திரன், ச. 2006. தமிழ்ச் சொல்வலை. [Tamil WordNet]. சித்திரபுத்திரன் மற்றும் பிறர் (பதிப்பு)ஆய்வுமுகங்கள். எச்.. ஆசிரியர் பேரவை. தமிழ்ப் பல்கலைக்கழகம். தஞ்சாவூர், 2006, 105-117.

இராஜேந்திரன், ச.எழுத்து மற்றும் இலக்கண திருத்தி. academia.edu

இராஜேந்திரன், ச. தமிழ்ச் சொல்வலை. academia.edu

இராஜேந்திரன், ச. விரிதரவு மொழியியல். academia.edu

இராஜேந்திரன், ச. இயந்திர மொழிபெயர்ப்பும் அதற்கான கருவிகள் உருவாக்கமும். academia.edu

இராஜேந்திரன், ச. தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வி. academia.edu

இராசேந்திரன் ச. கணினி மொழியியல் [Computational linguistics]. Uploaded in academia.edu and Research Gate.

இராசேந்திரன் ச. இயற்கைமொழி ஆய்வு [Natural language processing]. Uploaded in academia.edu and Research Gate.

இராசேந்திரன் ச. விரிதரவு மொழியியல் [Corpus linguistics]. Uploaded in Academia.edu

இராசேந்திரன் ச. தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வி [Morphological Analyzer for Tamil Verbs]. Uploaded in academia.edu and Research Gate.

இராசேந்திரன் ச. தமிழில் இயந்திர மொழிபெயர்ப்பும் அதற்கான கருவிகள் உருவாக்கமும் [Machine translation and Preparation of tools for machine translation]. Uploaded in academia.edu and Research Gate.

இராசேந்திரன் ச. தமிழில் எழுத்துப்பிழைத் திருத்தியும் இலக்கணப்பிழைத் திருத்தியும் [Spell and Grammar Checker for Tamil]. Uploaded in academia.edu

இராசேந்திரன் ச. சொல்வலை [Word net]. Uploaded in Academia.edu. Uploaded in academia.edu and Research Gate.

இராசேந்திரன் ச. 2019. தமிழில் ஒலியியக்கவியலும் உரையிலிருந்து பேச்ச்சாக்கமும் பேச்சிலிருந்து உரையாக்கமும் (Acoustic phonetics and text to speec processing and speech to text processin). Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:10 October 2019

இராசேந்திரன், ச. 2019. "தமிழில் சொற்பொருள்மயக்கநீக்கம் (Word sense disambiguation in Tamil)". Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:9 September 2019

இராசேந்திரன் ச. மற்றும் பாஸ்கரன், ச. 2006. தமிழ் மின்சொற்களஞ்சியம். [Tamil electronic Thesaurus) தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர். Uploaded in academia.edu. and Research Gate.

இராசேந்திரன் ச. மற்றும் க. அனிதா. 2019. தமிழ்ச் சொற்றொகையின் மூலப்பொருண்மையியல் ஆய்வு (Ontology of Tamil Vocabulary), Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:8 August 2019

சுந்தரம், இல. 2015. கணினித்தமிழ். விகடன் பிரசுரம்.

சங்க இலக்கியப் பொருட்களஞ்சியம். தமிழ்ப் பல்கலைக்கழகம்,தஞ்சாவூர்.

சுந்தரம், இல. 2015. கணினித்தமிழ். விகடன் பிரசுரம்.

சுப்பையா பிள்ளை, 2000. இயற்கை மொழி ஆய்வு. உலகத்தமிழாராய்ச்சி நிறுவனம், சென்னை.

செவ்வேள் கபிலன். 1994. கணிப்பொறிவழி தமிழ் வினைகளின் பகுப்பாய்வு. புத்தாக்க மொழியியியல் கழகம், சென்னை

பாஸ்கரன், ச. 2004. தமிழில் கணிப்பொறியியல், கணிப்பொறியில் தமிழ். உமா பதிப்பகம், தஞ்சாவூர்.

பெருஞ்சொல்லகராதி. தமிழிப் பல்கலைக்கழகம், தஞ்சாவூர்.

ராமகிருஷ்ணன், எஸ். (பதிப்பாசிரியர்) 2008.க்ரியாவின் தற்காலத்தமிழ் அகராதி. க்ரியா, சென்னை.

A Common Parts-of-Speech Tagset Framework for Indian Languages. Available from: https://www.researchgate.net/publication/220746407_A_Common_Parts-of-Speech_Tagset_Framework _for_Indian_Languages [accessed Dec 12 2019].

Abinaya, N., John, N., Ganesh, B. H., Kumar, A. M., and Soman, K. (2014). "Amrita cen@ fire-2014: Named entity recognition for Indian languages using rich features." In: Proceedings of the Forum for Information Retrieval Evaluation, pages 103–111. ACM.

Abney S. 1991. Parsing by Chunks. In: R. Berwick, S. Abney and C. Tenny (eds.) Principle–based Parsing . Kluwer Academic Publishers, Dordrecht, 1991.

Abney S. 1996. Partial Parsing via Finite-State Cascades. In: Proceedings of the ESSLLI'96 Robust Parsing Workshop, Prague, Czech Republic, 1996.

Aitchison, J. 1987. Words in the Mind: An Introduction to the Mental Lexicon. Oxford: Basil Blackwell.

Akila, A. & Chandra, E. 2013. "Isolated Tamil word speech recognitionsystem using HTK." International Journal of Computer Science Research and Application, 03/02, 30-38.

Akilan R. and Naganathan E. R. 2015. Morphological Analyzer For Classical Tamil Text: A Rule-Based Approach Vol. 10, No 20, November, 2015 ISSN 1819-6608, ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). www.arpnjournals.com 9325.

Akilandeswari A. and Sobha Lalitha Devi. 2015. "Tamil Pronominal Resolution Boosted By Sentence Transformation". Aust. J. Basic & Appl. Sci., 9(23): 566-572, 2015.

Akilandeswari A and Sobha Lalitha Devi (2014). "Anaphora Resolution in Tamil Novels", In Rajendra Prasath, Philip O'Reilly, T. Kathirvalavakumar (Eds), Mining Intelligence and Knowledge Exploration, Springer LNAI Vol 8891, pp. 268-277.

Akilandeswari A., Sobha Lalitha Devi. (2013). "Conditional Random Fields Based Pronominal Resolution in Tamil ", In International Journal on Computer Science and Engineering, Vol. 5 Issue 6 pp 601 – 610.

Akshar Bharati, Vineet Cahaitanya and Rajeev Sangal. 1995. Natural Language Processing. A Paninian Perspective. New Delhi: Prentice Hall of India.

Alex Graves, Santiago Fernandez, Faustino Gomez, and Jürgen Schmidhuber (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets. Proceedings of ICML'06, pp. 369–376.

Anandan P, Rajani Parthasarathy, Geetha, T.V. 2001. “Morphological analyzer for Tamil”. ICON 2002, RCILTS-Tamil Anna University, Chennai.

Anandan P, Ranjani Parthasarathi and Geetha T.V. 2001. Morphological Generator For Tamil Tamil Inayam, Malaysia 2001.

Anandan, P, Rajani Parthasarathy, Geetha, T.V. 2001. “Morphological Generator for Tamil” in Tamil Internet 2001 Conference Proceedings, Malaysia.

Ananth Ramakrishnan A. Sankar Kuppan and Sobha Lalitha Devi (2009) "Automatic Generation of Tamil Lyrics for Melodies." In: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity

Anbarasan, N. 2001. “Tamil Unicode-What do we need?”, paper read in National Seminar on Computational Linguistics and Dravidian Languages. CAS in Linguistics, Annamalai University, Annamalai.

Annamalai, E. “Corpora Development in Indian Languages”, in Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

Boulder, Colorado:Association for Computational Linguistics, Pages: 40–46

Ananth Ramakrishnan A, Sobha Lalitha. 2010. "An alternate approach towards meaningful lyric generation in Tamil". Proceedings of the Workshop: Second Workshop on Computational Approaches to Linguistic Creativity, June 5, Los Angeles, California, 2010, 31-39. Anand Kumar, M., Dhanalakshmi, V., Soman, K.P., Rajendran, S. (2009) A Novel Approach for Tamil Morphological Analyzer. In: Proceedings of Tamil Internet Conference 2009, Cologne, Germany, pp. 23–35 (October 2009).

Anand Kumar M, Dhanalakshmi V, Rekha R U,Soman K, P and Rajendran (2010). "A Novel data driven algorithm for Tamil morphological generator", IJCA,Vol.6,No.12,Pages:52-56.

Anand Kumar M, Dhanalakshmi V, Soman K.P , Rajendran S (2010) " A Sequence Labeling Approach to Morphological Analyzer for Tamil Language", IJCSE, Vol. 02, No. 06,, 1944-1951.

Anand Kumar M, Dhanalakshmi V, Soman K.P and Rajendran S (2010). "A Sequence Labelling Approach to Morphological Analyzer for Tamil Language". IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 1944-1951

Anand Kumar M, Dhanalakshmi V, Soman K P and Sharmiladevi V (2013). "Improving the Performance of English-Tamil Statistical Machine Translation System using Source-Side Pre-Processing", Proceedings of International Conference on Advances in Computer Science, AETACS © Elsevier, 2013

Anand Kumar M, Rajendran S. and Soman K. P. 2014 “AMRITA@ FIRE-2014: Morpheme Extraction for Tamil using Machine Learning (Working notes)”, in International Workshop: "MET shared Task" Forum for Information Retrieval Evaluation (FIRE- 2014), Bengaluru , 2014.

Anand Kumar M, Rajendran S and Soman K P (2014) “Tamil Word Sense Disambiguation using Support Vector Machines with rich features.” International Journal of Applied Engineering Research. ISSN 0973-4562, Volume X, Number X (2014) pp. xxx-xxx © Research India Publications http://www. ripublication.com/ijaer.htm

Anand Kumar, M., S. Rajendran, and K. P. Soman. 2014. “Supervised Cross-lingual Preposition Disambiguation for Machine Translation" iDravidian’ 2014. Symposium on Natural Language Processing for Dravidian Languages, ICON 2014.

Anand Kumar M, Rajendran S. and Soman K. P. 2015. “Cross-Lingual Preposition Disambiguation for Machine Translation.” Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015). Available online at www.sciencedirect.com

Anand Kumar M, Dhanalakshmi V, Soman K.P. and Rajendran S. (2014). “Factored Statistical Machine Translation System for English to Tamil Language.” Social Sciences & Humanities Journal homepage: ttp://www.pertanika.upm.edu.my/Pertanika J. Soc. Sci. & Hum. 22 (4): 1045 -1061 (2014)

Anand Kumar M. (2015). Tamil Linguistic Tools and English-Tamil Machine Translation System. United State: LAP Lambert Academic Publishing.

Anand Kumar M, S. Shriya, and K. P. Soman. (2015). AMRITA-CEN@FIRE 2015: Extracting entities for social media texts in Indian languages. CEUR Workshop Proceedings, 1587:85–88, 2015.

Anand Kumar M, Shivkaran Singh, Kavirajan B and Soman K P 2016. Shared Task on Detecting Paraphrases in Indian Languages (DPIL): An Overview. Anand Kumar M (http://orcid.org/0000-0003-0310-4510)

Anand Kumar M, Premjith B, Shivkaran Singh, Rajendran S and Soman K P. (2017). An Overview of the Shared Task on Machine Translation in Indian Languages (MTIL) - 2017 (Special issue on MTIL 2017) J. Intell. Syst. (), 1–16 DOI 10.1515/jisys-- © de Gruyter.

Anandan P., Saravanan K., Parthasarathy R., and Geetha T.V., (2002), "Morphological Analyzer for Tamil", in the proceedings of ICON 2002, Chennai.

Anderson, J.A. An Introduction to Neural Networks. MIT Press, Cambridge MA.

Anandan, P, Rajani Parthasarathy, Geetha, T.V. 2001. “Morphological Generator for Tamil”, in Proceedings of the Tamil Internet 2001 Conference.

Ananthapadmanabha, T. V. 1995. "Acoustic factors determining perceived voice quality," in Vocal fold Physiology - Voice quality control, O.Fujimura and M. Hirano, Eds. San Diego, Cal.: Singualr publishing group, 1995, ch. 7, pp. 113–126.

Ananthapadmanabha T V, A P Prathosh, A G Ramakrishnan. 2015. “Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index,” Journal of the Acoustical Society of America, Vol. 137, 2015.

Anbarasan, N. 2001. “Tamil Unicode-What do we need?”, paper read in National Seminar on Computational Linguistics and Dravidian Languages. CAS in Linguistics, Annamalai University, Annamalai.

Anderson, S.1988. Morphology as a parsing problem. Linguistics 26: 521-544.

Annamalai, E. “Corpora Development in Indian Languages”, in Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

ANUVADAKSH – An Expert English to Indian Language Machine Translation System.

Antony P J, Warrier N.J. and Soman K P. (2010). "Penn Treebank-Based Syntactic Parsers for South Dravidian Languages using a Machine Learning Approach", International Journal of Computer Applications (0975 – 8887) Volume 7– No.8, October 2010.

Aparna K G and A G Ramakrishnana. (2002). "Complete Tamil Optical Character Recognition System." 5th International Tamil Internet Conference (TIC 2002) at San Francisco, CA, USA

Ariaratnam I., Weerasinghe A. R. and Liyanage C. (2014). "A shallow parser for Tamil". 14th International Conference on Advances in ICT for Emerging Regions (ICTer), 10-13 Dec. 2014.

Aronoff, M. 1985. Word formation in Generative Grammar. Massachusetts: Masschusetts Institutue of Technology.

Aronoff, M. and Anshen F. 1998. Lexicalizatiion and productivity.” The Handbook of Morphology (ed) Andrew Spencer and Arnold M.Zwicky 237-247, Blackwell, Oxford.

Arulmozi S. 1998. Aspects of Inflectional Morphology – A computational Approach, Unpublished PhD Thesis. Hyderabad: University of Hyderabad.

Arulmozhi, P., Sobha, L., Kumara Shanmugam, B. 2004. Parts of Speech Tagger for Tamil, Symposium on Indian Morphology, Phonology & Language Egnineering, Indian Institute of Technology, Kharagpur.

Arulmozhi, P and Sobha, L. 2006. A Hybrid POS tagger for a Relatively Free Word Order Language. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages, pages 79-85.

Arulmozhi S. 1998. Aspects of Inflectional Morphophonology - A Computational Approach. Unpublished Ph.D Thesis. Hyderabad: University of Hyderabad.

Arulmozi, S and M.C. Kesava Murty. 2006. Verb Sense Disambiguation in Telugu-Tamil MT. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages pages 107-111.

Arulmozhi. P, Sobha. L. 2006. "Semantic Tagging for Language Processing". presented in 34th All India conference for Dravidian Linguistics held at International School of Dravidian Linguistics, Trivandrum.

Arulmozhi P, Sobha L and Kumara Shanmugam B. 2004. "Parts of Speech Tagger for Tamil", Symposium on Indian Morphology, Phonology & Language Engineering, March 19-21, IIT Kharagpur. : 55-57.

Arun, M.C. “Evaluation of Word Sense Disambiguation methods applied on Tamil Corpus. PhD thesis.

Arun, S., Anand Kumar, M., Soman, K. P. (2015) “Sentiment Analysis of Tamil movie reviews via Feature Frequency Count,” IJAER, 2015, pp.17934-17939.

Arun Srivatsa. R , Shriya.P , Adithya Srinivasan (2015) “Interpretation of Tamil Script using Optical Character Recognition in Android.” International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (5) , 2015, 4528-4530.

Arunselvan S.J., M Anand Kumar K. P. Soman (2015) "Sentiment analysis of Tamil movie reviews via feature frequency count" International Journal of Applied Engineering Research vol. 10 no. 20 pp. 17934-17939 2015.

Asher. 1995. Tamil. London: Croom Helm.

Atal, B.S.and Hanauer S. 1971. Speech analysis and synthesis by prediction of the speech wave. Jounal of Acoustical Society of America, 50: 637-655.

Attwell, E. and Elliott, S. 1987. ”Dealing with ill-formed English text.” In: The Computational analysis of English. Longman.

Aug 22, 2019 . The Unicode Standard: A Technical Introduction.

Balaji J, Geetha T V, Ranjani Parthasarathi. (2012). Semantic Parsing of Tamil Sentences. Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 15–22, COLING 2012, Mumbai, December 2012.

Banu M, Karthika C, Sundaramani P and Geetha T.V. 2007. "Tamil Document Summarization using Semantic Graph Method." International Conference on Computational Intelligence and Multimedia Applications." IEEE, pp 125-134, 2007.

Dravidian Linguistics, Thiruvananthapuram, India.

Balakrishnan, R. 2002. Morphology and Tamil Computing. Paper read in International Seminar on Tamil Computing, February 27, 28, 2002, Madras University.

Baskaran, S. “User Interface with Computers in Tamil-an Overview [kaNippoRi uraiyaakam]”, in Proceedings of the Third Tamil Nadu Science Congress, Pondicherry.

Baskaran, S. “Zip’s Law: A perspective study in Tamil”, in Proceedings of Eighth International Conference of Tamil Studies.

Baskaran, S. Report on Standardization of Tamil Key Board, Submitted to the Committee formed for Standardization of Tamil Key board by Tamilnadu Government.

Baskaran, S. “Tamil Computing”, tamizh vaLarcci.

Baskaran, S. “Evolution of Tamil computing”, Souvenir, National Symposium on Current Trends in Computer Applications.

Baskaran, S. “Experiment in Implementation of Finite-State-Machine Parser for Tamil Sentences”, in Computer Society of India Communications, Bombay.

Baskaran, S. “Compute as a Language Research Tools,” in Proceedings of the Second Tamilnadu Congress, Madras.

Baskaran, S. 2002. Semantic Analyser for Word Sense Disambiguation. MS Thesis, Anna University, Chennai.

Baskaran Sankaran, Vaidehi V (2002) "Role of Collocations and Case-Markers in Word Sense Disambiguation: A Clustering-Based Approach". In Proceedings of IEEE International Symposium on Natural Language Processing and Knowledge Engineering 2002. Vol. I. pp. 625-630. Hammamet, Tunisia.

Baskaran, S. 2003. “Word Sense Disambiguation in Tamil”, International Journal of Dravidian Linguistics.

Baskaran, S and Vaidehi, V. 2004. “Collocation Based Word Sense Disambiguation using Clustering for Tamil”, International Journal of Dravidian Linguistics 33.1: 13-28.

Baskaran, S. Vijay-Shanker, K. 2003. “Influence of Morphology in Word Sense Disambiguation for Tamil”, in Rajeev Sangal et al (eds.) Recent Advances in Natural Language Processing. Mysore: Central Institute of Indian Languages.

Baskaran, S et al. 2008. Designing a Common POS-Tagset Framework for Indian Languages. Paper presented in The 6thWorkshop on Asian Language Resources. Jauary, 2008, Hyderabad.

Baskaran, S. et al. 2007. Framework for a Common Parts-of-Speech Tagset for Indic Languages. Available in http://research.microsoft.com/~baskaran/POSTagset Cloeren, J. (1999) Tagsets. In Syntactic Wordclass Tagging, ed. Hans van Halteren, Dordrecht.: Kluwer Academic
.

Bellman P. 1957. Dynamic programming. Princeton University Press, Priceton, Princeton NJ.

Bessely K.R. and Karttunen L. 2000. Finite state Non-concatenative Morphotatics. Proceedings of the Fifth Workshop of ACL Special Interest Groups in Computational phonology.1-12, Luxemboug.

Becker Tilman and Rambow Owen.1994. 'Parsing Free Word-Order Languages in Polynomial Time', University of Paris.

Bergler, S. 1995. From lexical semantics to text analysis. In: Saint-Dizier, P and Viegas, E. Computational Linguistics. Cambridge Universtiy Press.

Bharati Akshar, Chaitanya Vineet and Sangal Rajeev .1995. 'Natural Language Processing', PHI.

Bharati Akshar and Sangal Rajeev. 1993. 'Parsing Free Word Order Languages in the Paninian Framework', ACL, Ohio, USA.

Bharati, A., Chaitanya, V., Kulkarni, A. P., & Sangal, R. 1997. ANUSAARAKA: Machine Translation in Stages. A Quarterly in Artificial Intelligence, 10(3), 22-25.

Bharati Akshar, Sangal Rajeev and Reddy T. Papi .2002. 'A Constraint Based Parser Using Integer Programming', International conference on NLP, Mumbai, India.

Barathi Ganesh HB, Anand Kumar M and Soman KP. "Conditional Random Fields for Code Mixed Entity Recognition."

Beesley, K.R. and L.Karttunen. 2003. Finite-state Morphology. Plato Alto, Calif:CSLI Publications.

Black, H. A. Black, A.C. 1999. A Conceptual introduction to morphological parsing using AMPLE. Dallas:Summer Institute of Linguistics.

Brochure on ‘Language Technology Products’ of the Resource Centre for Indian Language Technology Solutions-Tamil, Chennai.

Calder, J. 1989. “Paradigmatic Morphology”. Proceedings of the 4th Meeting of the European Chapter of the Association for Computational Linguistics (ACL’89) (Manchester), 58-65.

Caramazza A. Laudanna A and Romani C. 1988. Lexical access and inflectional morphology. Cognition 28:297-332.

Caramazza A, Hillis AE, Rapp BC, Romani C .1990. The multiple semantics hypothesis: multiple confusions? Cogn Neuropsychol 7:161–189

Caramazza, A. and Shelton, J.R. 1998. Domain-specific knowledge systems in the brain: the animate-inanimate distinction. J. Cogn.Neurosci. 10, 1–34

Carstairs-McCarthy A. 1998. Paradigmatic structure: Inflectional paradigm and morphological classes. In: Specer A and Zwicky A. (eds). Handbook of morphology. Oxford: Blackwell.

Chandrakanth, D. Anand Kumar M. and Gunasekaran S. (2012) "Parts-of-Speech tagging for Tamil language‖", IJCSE,Vol.6,No.6,Pages:88-93.

Chandrasekar, M. & Ponnavaikko, M. (2008). Tamil speech recognition: A complete model. Electronic Journal Technical Acoustics (EJTA).

Chandrasekaran S, Chithra S. (2016) Big Data Methods for Computational Tamil Linguistics with Emotion Oriented Semantic Pattern Mining. Discovery, 2016, 52(245), 979-984.

Charles, A. P. H. & Devaraj, G. 2004. Alaigal-A Tamil speech recognition. Tamil Internet, Singapore.

Chellamuthu, K.C. et al. 1984. Tamil University Machine Translation System (TUMTS), Thanjavur: Tamil University.

Chellamuthu, K.C. 2002. 'Russian to Tamil Machine Translation System at Tamil University,' in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, 74-83.

Chellamuthu, KC. 2019. Russian to Tamil Machine Translation System at Tamil University. (Down loaded from internet on 23.08.2019.)

Chithiraputhiran, H. Tamil Lexical Resource: An electronic Data Base. Paper read in International Seminar on Tamil Computing, February 27, 28. Madras University.

Church, K. W. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In: Proceedings of the 1st Conference on Applied Natural Language Processing, ANLP, pages 136–143. ACL, 1988.

Clark, John; & Yallop, Colin. 1995. An introduction to phonetics and phonology (2nd ed.). Oxford: Blackwell. ISBN 0-631-19452-5.

Cruse, D.A. 1986. Lexical Semantics. Cambridge: Cambridge University Press

Cruz 2002. Parsing and generation of Tamil verbsin GSMorph. Unpublished MPhil Dissertation. Hyderabad: University of Hyderabad.

Dameraru, F.J.1964. “A technique for computer detection and correction of spelling errors”. Communications of the ACM, 7(3), 171-176.

Denial Naber et al

Deivasundaram, N. and Gopal, A. 2003. ‘Computational Morphology of Tamil’ In: B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, 406-410.

Database for Indian Languages□, Speech and Vision Lab, IIT Madras, India, 2001.

Dhanalakshmi V, Anandkumar M, Vijaya M.S, Loganathan R, Soman K.P. and Rajendran S. (2008). "Tamil Part-of-Speech tagger based on SVMTool", In: Proceedings of the COLIPS International Conference on natural language processing(IALP), Chiang Mai, Thailand. 2008:59-64.

Dhanalakshmi V, Anandkumar M, Shivapratap G, Soman, K P and Rajendran S. (2009). Tamil POS Tagging using Linear Programming, In: International Journal of Recent Trends in Engineering, 1(2):166-169.

Dhanalakshmi V, Anand kumar M, Rajendran S and Soman K P. (2009). "POS Tagger and Chunker for Tamil Language," Proceedings of the 8th Tamil Internet Conference. Cologne, Germany, 2009.

Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P., and Rajendran S. (2009). “Chunker For Tamil Using Machine Learning” 7th International Conference on Natural Language Processing 2009 (ICON2009), IIIT Hyderabad, India, December 2009.

Dhanalakshmi V, Padmavathy P, Anand Kumar M, Soman K P, and Rajendran S 2009, “Chunker for Tamil”, Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, IEEE Press, doi: 10.1109/ARTCom. 2009.191.

Dhanalakshmi V. and Anand Kumar M. “Hierarchal POS tagging for Tamil language using Machine learning approach”, Unpublished. Uploaded in academia. edu.

Dhanalakshmi V, Anand Kumar M, Shivapratap G, Soman K.P and Rajendran S (2009) "Tamil POS Tagging using Linear Programming", International Journal of Recent Trends in Engineering, Vol. 1, No. 2.

Dhanalakshmi V, Anand kumar M, Rajendran S, Soman K P. (2009). POS Tagger and Chunker for Tamil Language.

Dhanalakshmi V., Anand Kumar M., Rekha R.U., Arun Kumar C., Soman K.P. and Rajendran S. 2009. "Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches," artcom, pp.433-435, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009,IEEE Press,doi: 10.1109/ARTCom.2009.184.

Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P. and Rajendran S. 2009. "Chunker for Tamil," artcom, pp.436-438, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009, IEEE Press,doi: 10.1109/ARTCom.2009.191.

Dhanalakshmi, V., Anand Kumar, M., Rekha, R.U., Soman, K.P., Rajendran, S. 2010. "Grammar Teaching Tools for Tamil Language", In: Technology for Education Conference (T4E 2010), pp. 85–88, India, 2010.

Dhanalakshmi V. 2010. Shallow Parser for Tamil. Thesis Submitted to Tamil University for the Award of the Degree of Doctor of Philosophy in Linguistics in Department of Linguistics, Tamil University, Thanjavur.

Dhanalakshmi V and Anand Kumar M, 2011. "Tamil shallow parser using machine leaning Approach", Tamil Internet.

Dhivya, R, Dhanalakshmi V, Anand Kumar M and Soman K.P. 2011. “Clause Boundary Identification for Tamil Language Using Dependency Parsing.” SPIT/IPC 2011: 195-197.

Dhurai Pandi, 2002. “Morphological Generator and Parsing Engine for Tamil Verb Forms (abstract)”, in Kalyanasundaram (ed) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 59.

Durai Pandi. 2002. “English-Tamil Machine Translation System”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 86.

Durai Pandi, 2002. “Morphophonemic Rules for Tamil Computing”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 98.

Duraipandi, R. 2006. “The Mophological Generator and Parsing Engines of Tamil Verb forms”. In Tamil Internet 2006.Chennai: Asian Printers.

Dowty, D. 1991. “Thematic proto-roles and Argument selection”. Language vol 67-3.

EAGELES Recommendations for the Morphostyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R Version of Mar, 1996.

EEN540-Project1- The Acoustic Features of Speech Sounds & Frequency Domain Filtering.

Elangovan, A. “Optimisation Techniques in Unicode Tamil Font Development”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 30.

Fant, Gunnar. 1960. Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations. Description and analysis of contemporary standard Russian (No. 2). s'Gravenhage: Mouton. (2nd ed. published in 1970).

Fellbaum, C. 1990. 'English verbs as a semantic net.' In: International Journal of Lexicography 3 (4):278 - 301.

Fellbaum, C. 1993. 'English Verbs as a Semantic Net.' (Down loaded from internet)

Fellbaum, C. (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press

Fellbaum, C., Gross, D. and Miller, K. 1993. 'Adjectives in WordNet.' (Down loaded form internet).

Cook, Walter A., SJ .1989. Case Grammar Theory. Washington, DC: Georgetown University Press.

Fillmore, Charles J. 1968. "The Case for Case". In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88.

Final Report: Development of Corpora of Texts of Indian Languages in Machine Readable form: Part II (Tamil, Telugu, Kannada, Malayalam), TDIL-Corpora Group, CIIL, Mysore January 1995.

Flanagan, James L. (1972). Speech analysis, synthesis, and perception (2nd ed.). Berlin: Springer-Verlag. ISBN 0-387-05561-4.

Francis Ekka and Ganesan, M. 1994. “Issues on Standard ISCII Codes and Inscript Keyboard”, Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

Frost, R., Katz, L., & Bentin, S. .1987.. Strategies for visual word recognition and orthographical depth: A multilingual comparison. Journal of Experimental Psychology: Human Perception and Performance.

Forster, K. I. 1976. Accessing the mental lexicon. In R.J.Wales & E.Walker (Eds.), New Approaches to Language Mechanisms. (pp. 257-287). Amsterdam: North-Holland.

'The Unicode Standard: A Technical Introduction' 8/05/2001.

Ganesan, M. 1994. "Functions of the Morphological Analyser Developed at CIIL, Mysore", in Harikumar Basi (ed.) Automatic Automatic Translation (seminar proceedings), Thiruvananthapuram: ISDL.

Ganesan, M. 1994. "'A Scheme for Grammatical Tagging of Corpora of Indian Languages", in B.B. Rajpurahit (ed.) Technology and Languages, Mysore:CIIL

Ganesan, M. 1996. "Relevance on Indian Grammatical Theories for Automatic Translation: chances and challenges", paper presented in National Seminar of Dravidian Linguists Association, Dravidian University, Kuppam, June 1996.

Ganesan, M. 2000. "Compilation of Electronic Dictionary for Tamil", Annamalai University, Annamalai Nagar. www.bhasshaindia.cjb.net/G7ganesan.pdf

Ganesan, M. 2003. "Computational Morphology of Tamil", in B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, pages 399-405.

Ganesan, M and Francis Ekka. 1994. "Morphological analyzer for Indian Languages", Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, Newdelhi: BPB Publication.

Garman, M. 1990. Psycholinguistics. Great Britain: Cambridge University Press.

Ganesh, J., Parthasarathi, R., Geetha, T. V., and Balaji, J. (2014). Pattern based bootstrapping technique for Tamil POS tagging. In: Mining Intelligence and Knowledge Exploration, Lecture Notes in Computer Science, pages 256–267. Springer, Cham.

Girish Nath Jha, The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI)., in: LREC, 2010.

Grainger J., Cole P. and Segui J. 1991. Masked morphological priming in visual word recognition. Journal of memory and Language 30: 370-384.

Green, B and Rubin, G. 1971. Automated Grammatical Tagging of English. Department of Linguistics, Brown University, 1971.

Grimshaw, J. 1990. Argument structure. MIT Press.

Gowri Mani G. Latent Semantic Analysis and Applications for Tamil Documents, MCA Project, Anna University, Chennai.

Halle M. 1973. “Prolegomena to a Theory of Word Formation.” Linguistic Inquiry Volume 4 Number i (Winter, I973) 3-I6.

Harel, D. 1987. Algorithmics: The sprit of computing. Great Britain: Wesley Publishers.

Hardcastle, William J.; & Laver, John (Eds.). (1997). The handbook of phonetic sciences. Oxford: Blackwell Publishers. ISBN 0-631-18848-7.

Haşim Sak, Andrew Senior, Kanishka Rao, Françoise Beaufays and Johan Schalkwyk (September 2015): "Google voice search: faster and more accurate." Archived 9 March 2016 at the Wayback Machine.

Hewavitharana, S. 2002. ‘A Two Stage Classification Approach to Tamil Handwriting Recognition’ in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 118-124.

Heylen, D. 1995. “Lexical functions, generative lexicons and the world”. In Saint-Dizier, P and Viegas, E. (eds). Computational lexical semantics. Cambridge University Press.

Holmes, V.M. and O'Regan J.K. 1992. Reading derivationally affixed french words. Language and Cognitive Processes, 7 (1992), pp. 163-192

Hooper, B.J. 1976. An Introduction to Natural Generative Phonology. New York: Academic Press.

Ide, N. and V´Eronis, J. 1998. “Word sense disambiguation: The state of the art.” Computat. Ling. 24, 1, 1–40.

Information Technology Department. ‘Information Technology – Standardization of Tamil Key Board and Encoding of Tamil. Glyphs – Recommendations of the Sub-committee on Tamil in Information Technology – Accepted – orders – Issued. (G.O. Ms.NO.17 dated” 13 June 1999).

Jackendoff, R. 1987. “The status of thematic role in linguistic theory”. Linguistic Inquiry 18: 369-411.

Jackendoff, R. 1988. Consciousness and computational mind. MIT Press.

Jackendoff, R. 1990. Semantic structure. MIT Press.

Jakobson, Roman; Fant, Gunnar; & Halle, Morris. 1952. Preliminaries to speech analysis: The distinctive features and their correlates. MIT acoustics laboratory technical report (No. 13). Cambridge, MA: MIT.

Jalin A.F. and J. Jayakumari (2017) "Text to speech synthesis system for Tamil using HMM," (2017) IEEE International Conference on Circuits and Systems (ICCS), Thiruvananthapuram, 2017, pp. 447-451. doi: 10.1109/ICCS1.2017.8326040.

Janarthanam, S., Nallasamy, U., Ramasamy, L., Santhoshkumar, C. (2007). Robust Dependency Parser for Natural Language Dialog Systems in Tamil, In Proceedings of 5th Workshop on Knowledge and Reasoning in Practical Dialogue Systems(IJCAI KRPDS-2007), pp. 1–6, Hyderabad, India, 2007.

Jayaram, B.D. “Development of Corpora in Indian Languages”, paper presented in the Seminar in the use of Computers in Indian Languages, held in CIIL, Mysore, August 1992.

Jayavardhana Rama, G.L. Ramakrishnan, A.G., Muralishankar R. and Prathibha P. (2002). "A Complete Text-to-Speech Synthesis System in Tamil," Proc. IEEE 2002 Workshop on Speech Synthesis, Sep. 11-13, Santa Monica, CA USA, 2002.

Jean-Francois Le Ny, 1995. “Mental lexicon and machine lexicon: which properties are shared by machine and mental word representation? Which are not?”. In: Saint-Dizier P and Viegas, E. 1(eds) 1995. Computational lexical semantics. Cambridge Universiy press.

Jensan 1993

Jensen, K, Heidorn,G.E., Richardson S.D. (eds.) 1993. Natural language Processing : The PLNLP approach, Kluwer Academic Publishers.

Johnson, Keith (2003). Acoustic and Auditory Phonetics (Illustrated). 2nd edition by Blackwell Publishing Ltd. ISBN 1-4051-0122-9 (hardback: alkaline paper); ISBN 1-4051-0123-7 (paperback: alkaline paper).

Jurafksy, D and Martin J.H. 2000. Speech and language processing: An Introduction to Natural Language Processing, Computational Linsutics, and Speech Recognition. Person Education, New Delhi.

Kalyanasundaram, K.A. “Comparison of transliteration schemes and key mapping of Tamil fonts”, http://www.Geocities.com/Athens/5180/translit.html 19/12/2000.

Kamakshi, S. 2002. “Preliminaries for Digitizing the Personal Pronouns in English into Tamil (Distribution Sensitive Machine Translation Aid–DSMTA)–A Demo Paper”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 87-97.

Kamakshi, S. 2004. “Machine Recognition and Translation of ‘~ing’ words in English into Tamil through bilingual Machine Tractable Dictionary”, paper presented in International Conference on Indian Lexicography 28-30 January 2004, CASL, Annamalai University, Annamalainagar.

Kamakshi, S. and Rajendran, S. 2004. Preliminaries to the preparation of a machine aid to translate Linguistics Texts written in English into Tamil. DLA Publications, Thiruvanathapuram.

Kamakshi Prasad V.., T. Nagarajan., Hema A. Murthy. 2004. ”Automatic segmentation of continuous speech using minimum phase group delay functions”, Speech Communications, Vol 42, pp 429-446, 2004.

Karttunen L., Kaplan M.R.and Zaenen A. 1992. Two-level Morphology with Composition. Xerox Palo Alto Research Centre, Stanfond University.

Kasirao, V. “Impact of Information Technology on Information Management Sevices with Special Reference to Tamil Computing: A Study”, paper read in International Seminar on Tamil Computing, February 27, 28. Madras University.

Kashket Michael B .1986. 'Parsing a Free-Word Order Language: Warlpiri', 24th proceedings of the ACL, New York, USA.

Katz L., Boyce S., Goldstein L and Lukatela G. 1987. Grammatical information effects in auditory word recognition. Cognition 25:235-263.

Kausika.N and V. Uma 2016. “Sentiment Analysis of English and Tamil Tweets using Path Length Similarity based Word Sense Disambiguation.” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 82-89.

Kavirajan, B & Kumar, M & Kp, Soman & Sankaravelayuthan, Rajendran & Vaithehi, Vaithehi. (2017). Improving the rule based machine translation system using sentence simplification (English to Tamil). 957-963. 10.1109/ICACCI.2017.8125965.

Keerthana S., Dhanalakshmi V., Anand Kumar M., Ajith V.P., Soman K.P. 2012. Tamil to Hindi Machine Transliteration Using Support Vector Machines. In: Das V.V., Ariwa E., Rahayu S.B.

(eds) Signal Processing and Information Technology. SPIT 2011. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 62. Springer, Berlin, Heidelberg

Kernighan, M.D., Church, K.W. and Gale,W.A. 1990."A spelling correction program base on a nosiy channel model. In COLING-90, Helsinki, vol II, 205-211.

Key, M. 1987. "Non-concatenative finite-state morphology". Proceedings of the 3rd Meetings of the European Chapter of the Computational Linguistics (ACL'96) Los Altols, Calif.), 159-166.

Kiran R., Nivedha K., Pavithra Devi S. and Subha T. (2017). Voice and speech recognition in Tamil language. 2017 2nd International Conference on Computing and Communications Technologies (ICCCT).

Klein, S and Simmons, R.F. 1963. A Grammatical coding of English words. JACM 10:334-347, 1963.

Knuth, D. E. 1973. The Art of Computer Programming: Fundamental algorithms, Volumes 1-2. Addison-Wesley.

Kohilavani S, T. Mala, T. V. Geetha. (2009). "Automatic Tamil Content Generation." 2009 International Conference on Intelligent Agent & Multi-Agent Systems.

Koskenniemi, K. 1984. "A general computational model for word-form recognition and production". Proceedings of the 10th International Conference on Computational Linguistics (COLING'84) (Stanford,Calif.)178-181.

Krishnamoorthy, V. "OCR Software for Printed Tamil Text", in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 99-101.

Krishnamoorthy, V. Challenges faced in developing Handwriting Recognition Software for Tamil. Downloaded from internet.

Kukich, K. 1992. "Techniques for automatically correcting words in text". ACM Computing Surveys, 24(4), 377-439.

Kumar A M, Rajendran, S, and Soman K P (2014). "Tamil word sense disambiguation using support vector machines with rich features", International Journal of Applied Engineering Research, vol. 9, pp. 7609-7620, 2014.

Kumar C.S. and Foo Say Wei (2003), 'A Bilingual Speech Recognition System for English and Tamil', Proceedings of Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia, Vol. 3, pp. 1641-1644.

Kumara Shanmugam, B. 2001. "Parse Representation of Tamil Syntax", Chennai.

Kumara Shanmugam, B. 2002. "Machine Translation as related to Tamil", in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 84-85.

Kumara Shanmugam B. 2002. 'Machine aided translation involving Tamil', International Conference on South Asian Languages (ICOSAL), Annamalai University, Chidambaram.

Kumara Shanmugam, B. 2004. Syntactic Parser for Tamil. M.S. (in NLP) dissertation submitted to Anna University, Chennai.

Kumara Shanmugam, P. 2004. Parse Representation of Tamil syntax. MS Thesis, Anna University,

Kumara Shanmugam B, 2004, 'Parsing a free word order language: Tamil', Simple'04, IIT, Kharagpur.

Kumara Shanmugam B and Ramanan S.V, 'Computational Framework for Parsing Tamil Sentences', International Journal of Dravidian Linguistics.

Ladefoged, Peter. 1996. Elements of Acoustic Phonetics (2nd ed.). The University of Chicago Press, Ltd. London. ISBN 0-226-46763-5 (cloth); ISBN 0-226-46764-3 (paper).

Lakshmana Pandian S. and T. V. 2008. Geetha Morpheme based Language Model for Tamil Part-of-Speech Tagging. (Downloaded from the website Research Gate).

Lakshmana Pandian, S. Geetha, T. V. 2018. Tamil Question Classification Using Morpheme Features. GoTAL '08 Proceedings of the 6th international conference on Advances in Natural Language Processing, Pages 272 – 283 ages 272 - 283 , Gothenburg, Sweden — August 25 - 27, 2008, Springer-Verlag Berlin, Heidelberg ©2008

Lakshmi A. and Hema A. Murthy. 2006. □"A Syllable based continuous speech recognizer for Tamil", Interspeech 2006, ICSLP, Pittsburgh, Sep. 2006.

Lakshmi S, Sindhuja Gopalan, Sobha Lalitha Devi. 2014. “Cross Linguistic Variations in Discourse Relations among Indian Languages.” In Proceedings of International Conference on Asian Language Processing 2014, Kuching, Sarawak, Malaysia.

Lakshmi S., and Sobha Lalitha Devi. 2014.”Rule-based Case Transfer in Tamil Malayalam MT”, In proceedings of CICLING, 2014.

Leavitt, R.R.J. MORPHE: A Practical Compiler for reversible morphology rules. Pittsburgh: Carnegie Mellon University.

Laudanna, A., Badecker, W., & Caramazza, A. 1992. Processing inflectional and deriva- tional morphology. Journal of Memory and Language, 31,333-348.

Laudanna, A., Cermele, A., & Caramazza, A. 1992. Morpholexical representations and reading. Paper presented at the Fifth Conference of the European Sociery for Cognitive Psychology, Paris, September.

Leech, G and Wilson, A. 1996. Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R.

Leech, G and Wilson, A. 1999. Standards for Tag-sets. In Syntactic Wordclass Tagging, ed. Hans van Halteren, Dordrecht: Kluwer Academic.

Lehmann, T. 1989. A grammar of Modern Tamil. Pondicherry: Pondicherry Institute of Linguistics and Culture.

Levenshtein, V.I. 1966. ”Binary code capable of correcting deletions, insertions and reversals“. Cybernetics and Control theory, 10(8), 707-710.

Loganathan, R. (2010). English-Tamil Machine Translation System. Master of Science by Research Thesis, Amrita Vishwa Vidyapeetham, Coimbatore.

Loganathan Ramasamy and ZdenZabokrtsky , (2011), "Tamil Dependency Parsing: Results Using Rule Based and Corpus Based Approaches", Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science Volume 6608, pp 82-95.

Loganathan, Ramasamy, OndrejBojar, Zdenek Žabokrtský, L. O. (2012), Morphological Processing for English-Tamil Statistical Machine Translation. In 24th International Conference on Computational Linguistics (p. 113).

Loganathan, Ramasamy, OndrejBojar, Zdenek Žabokrtský, L. O. (2012), Morphological Processing for English-Tamil Statistical Machine Translation. In 24th International Conference on Computational Linguistics (p. 113).

Lukatela, G. Gligorijerc, B., Kostic, A. Turvey M. 1980. Representation of inflected nouns in the interanal lexicon. Memory and Cognition 8:415-423.

Lyons, J.1977. Semantics, volume 1, New York: Cambridge University Press.

Lushanthan S, A. R. Weerasinghe, D.L. Herath. (2014) "Morphological Analyzer and Generator for Tamil Language" 2014 International Conference on Advances in ICT for Emerging Regions (ICTer).

Madhupriya U. Enhanced Version of Morphological Analyzer and Parser for Tamil Language. MCA Project, Anna University, Chennai.

Mahalakshmi, S., Anand Kumar, M., Soman, K.P., 2015. Paraphrase detection for Tamil language using Deep learning algorithm. International journal of Applied Engineering Research, 10 (17), pp. 13929-13934

Malarkodi C.S., Elisabeth Lex and Sobha Lalitha Devi. (2016). "Named Entity Recognition for the Agricultural Domain", In proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey. Special issue: Research in Computing Science.

Malarkodi C.S. and Sobha Lalitha Devi. (2018). "Twitter Named Entity Recognition for Indian Languages.", In proceedings of 18th International Conference on Computational Linguistics and Intelligent Text Processing, March 18 to 24, 2018, Hanoi, Vietnam.

Manone, A. G., S. Saravanan, R. Loganathan, and K. P. Soman. 2009. “Amrita Morph Analyzer and Generator for Tamil:A Rule-Based Approach.” Proceedings of Tamil Internet Conference 2009, Cologne, Germany, October 2009.

Manone, V.K., Rajendran S Anand Kumar M and Soman K P. A new TAG Formalism for Tamil and Parser Analytics. Uploaded in academia. edu and Research Gate.

Manone V.K., Rajendran S., Anand Kumar, M., and Soman, K. P., 2014. “A new TAG Formalism for Tamil and Parser Analytics”, iin iDravidian, International Symposium for Dravidian Languages, Co-located with ICON2014, Goa, 2014.

Manone, V.K. Rajendran S., & Soman, K.P. 2015. “A Synchronous Syntax for English-Tamil language pair for Machine Translation”. Fourth International Symposium on Natural Language Processing (NLP’15). Kochi, Kerala: Co-affiliated with Fourth International Conference in Computing, Communications and Informatics (ICACCI-2015).

Marimuthu K., Amudha K., Bakiyavathi T. and Sobha Lalitha Devi. (2013). ”Word Boundary Identifier as a Catalyzer and Performance Booster for Tamil Morphological Analyzer”, In Proceedings of 6th Language and Technology Conference, Human Language Technologies as a challenge for Computer Science and Linguistics - 2013, Poznan, Poland.

Marc Cavazza and Pierre Zweigenbaum. 1995. Saint-Dizier, P and Viegas, E. 1995. Computational lexical semantics. Cambridge University Press.

Marslen-Wilson, W. D. 1987. Parallel processing in spoken word recognition. Cognition 25: 71-102.

Marslen-Wilson, W., L. K. Tyler, R. Waksier, and L. Older. 1994. Morphology and meaning in me English mental lexicon. Psychological Review 101:3-33. Marslen-Wilson, W. and Welsh, A. 1978. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 10. 29–63.

Maxwell, M. 1998. Two Theories of Morphology, one implementation. Dallas: Summer Institute of Linguistics.

McConnel, S. and Black, H.A. 2002. AMPLE: reference manual:A morphological Parser for Linguistic Exploration; version 3.5. Dallas: Summer Institute of Linguistics.

McQueen, J.M., & Cutler, A. (1998). Spotting (different types of) words in (different types of) context. In R.A. Mannell & J. Robert-Ribes (Eds.), Proceedings of the 5th International Conference on Spoken Language Processing (vol. 6, pp. 2791–2794). Sydney: Australian Speech Science and Technology Association.

Meenakshi Narayanaswamy, Ravikumar.K.E., and K.Vijay-Shanker “A Biological Named Entity Recognizer “ accepted in Proceedings of the Pacific Symposium on Biocomputing '03 (PSB'03), Hawaii, January 2003.

Melanie Pinola. 2011. "Speech Recognition Through the Decades: How We Ended Up With Siri". PC World. Retrieved from internet.

Mel'čuk, I.A. 1988. Dependency syntax: Theory and practice. State University of New York Press.

Michael S. Kapalan. 2002. "Unicode and Tamil", in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 1.

Menaka S, Vijay Sundar Ram And Sobha Lalitha Devi. 2018. Morphological Generator for Tamil. https://www.researchgate.net/publication/265866650.

Menaka S., Pattabhi R. K. Rao, Sobha Lalitha Devi, (2011), "Automatic Identification of Cause-Effect Relations in Tamil Using CRFs", In A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol. 6608/2011 pp 316-327.

Menaka Sankarlingam, Malarkodi C S and Sobha Lalitha Devi (2014), "A Deep Study on Causal Relations and its Automatic Identification in Tamil", In proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation, Organized under LREC2014, Reykjavik, Iceland.

Michael S. Kapalan. 2002. "Unicode and Tamil", in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page

Miller, G.A.1990. 'Nouns in WordNet: a lexical inheritance system.' In: International Journal of Lexicography 3 (4): 245 - 264.

Miller, G.A. 1991. Science of Words. New York: Scientific American Library.

Miller, G.A. 1993. 'Nouns in WordNet: A Lexical Inheritence System.' (Down loaded from internet)

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.J.1990. 'Introduction to WordNet: an on-line lexical database.' In: International Journal of Lexicography 3 (4):235 - 244.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D, and Miller, K. 1993. 'Introduction to WordNet: An On-line Lexical Database.' (Down loaded from internet).

Mohan Raj, S.N. and Rajendran S. 2016. "Tamil Oriented Machine Translation under Indian Language to Indian Language Machine Translation (ILILMT) consortium." In: Proceedings of 15th World Tamil Internet conference 2016, held at Gandhi Gram Rural University, Tamil Nadu, September 8-11, 2016, pages 393-402.

Mokanarangan T, T Pranavan, U Megala, N Nilusija, Gihan Dias, Sanath Jayasena, Surangika Ranathunga. 2016. "Tamil morphological analyzer using support vector machines."

International Conference on Applications of Natural Language to Information Systems, 15-23, Springer, Cham

Muralidharan V. and Sharma D. M. (2016). Construction Grammar based Annotation Framework for parsing. Tamil. In 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey, Report No: IIIT/TR/2016/-1

Morton, J. 1968. A preliminary functional model for language behaviour. International Audiology, 1964, 3, 216–225. (Reprinted in R.C. Oldfield & J.C. Marshall (Eds.), Language. London: Penguin.

Morton, J. 1968. Grammar and computation in language behavior. In J.C. Catford (Ed.), Studies in language and language behavior. Center for Research in Language and Language Behavior Progress Report No. VI, University of Michigan.

Morton, J. 1969. The interaction of information in word recognition. Psychological Review, 1969, 76, 165–178.

Morton, J. 1978. Word recognition. In J. Morton and J.C. Marshall (Eds.), Psycholinguistics Series II, London: Elek Scientific Books.

Muthu Nedumaran. 2002.”Glyph Choices and Techniques for Building Unicode Based Tamil Fonts”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 31.

Nagarajan, T., H. A. Murthy and R. M. Hegde. 2003.□”Segmentation of speech in to syllable-like units”, in proceedings of EUROSPEECH, 2003, pp. 2893-2896.

Nagarajan, T. and H. A. Murthy. 2004.□”Group delay based segmentation of spontaneous speech into syllable-like units”, EURASIP Journal of Applied signal processing, vol. 17, pp.2614-2625, 2004.

Nagarajan T., Vijayalakshmi .P., and Shanina A. Development of Text-to-speech synthesis for Indian languages (Project funded by MCIT from 2012-17), SSN College, Chennai.

Nagarajan T. and Vijayalakshmi P. HMM-based Text-to-speech synthesis system for Malaysian Tamil. (Funded by Murasu systems Sdn Bhd, Malaysia 2016-2017), SSN College, Chennai.

Nagarajan, T. Vijayalakshmi P., Bharathi B., and Sasirekha. Speech enabled interactive enquiry system in Tamil. (Project funded by Tamil Virtual Academy, Chennai from 2016- 2017), SSN College, Chennai

Narayana Murthi, K.N, Sobha, L, Muthukumari, B (2007) "Pronominal Resolution in Tamil Using Machine Learning Approach" The First Workshop on Anaphora Resolution (WAR I),Ed Christer Johansson, Cambridge Scholars Publishing, 15 Angerton Gardens, Newcastle, NE5 2JA, UK pp.39-50.

Natanasabapathy, S. 2002. “Study of Segmental Duration of Tami sounds”, paper presented in International Seminar on Tamil Computing, 27-28 February and 1 March, 2002, University of Madras, Chennai.

Nathan Green, Loganathan Ramasamy and Zdenekˇ Zabokrtsk ˇ y´, (2012). "Using an SVM Ensemble System for Improved Tamil Dependency Parsing",Association for Computational Linguistics, pages:72–77.

Navaneethan, P et al. 2004. “Multilingual Word Processor with Online Noun Translator”, paper read in National Seminar on Cross Language Communication Tools (Machine Translation) for Indian Languages, February 25-27, 2004.

Navigli, R. 2009. Word Sense Disambiguation: A Survey. ACM Computing Surveys, Vol 41, No.2, Article 10, Publication date February 2009.

Navaneethan, P et al. 2004. “Multilingual Word Processor with Online Noun Translator”, paper read in National Seminar on Cross Language Communication Tools (Machine Translation) for Indian Languages, February 25-27, 2004.

Nayeemulla Khan A. and Yegnanarayana B. 2001. ‘Development of Speech Recognition System for Tamil for Small Restricted Task’, Proceedings of National Conference on Communication, India.

Nayeemulla Khan, A, Suryakanth, V, Gangashetty, & Yegnanarayana, B. 2003. “Syllabic Properties of Thee Indian Languages: Implications for Speech Recognition and Language Identification”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Naber, D. 2003. A Rule-Based Style and Grammar Checker. Diploma Thesis, Computer Science – Applied, University of Bielefeld.

Nida, E.A. 1975.a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton.

Nida, E.A. 1975.b. Exploring Semantic Structure. The Hague: Mouton.

Oflazer, K. 1999. Dependency Parsing with Extended finite state approach. Computing research Laboratory. New Mexico: New Mexico State University.

Ogden, R. 2009. An Introduction to English Phonetics. Oxford: Oxford Universiy Press.

Pisoni, David B.; & Remez, Robert E. (Eds.). (2004). The handbook of speech perception. Oxford: Blackwell. ISBN 0-631-22927-2.

Padmamala R. and V. Prema. 2017. “Sentiment analysis of online Tamil contents using recursive neural network models approach for Tamil language.” 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)

Pandian, S. L. and Geetha, T. V. 2009. CRF models for Tamil part of speech tagging and chunking. In: Proceedings of the 22Nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy, ICCPOL ’09, pages 11–22, Berlin, Heidelberg. Springer-Verlag.

Pattabhi R K Rao T and Sobha L 2008. "Identifying Similar and Co-referring Documents Across Languages", In Proceedings of 3rd International Joint Conference on Natural Language Processing Workshop on Cross Lingual Information Access (CLIA- 2008), Hyderabad, India. pp. 10 -17.

Pattabhi R K Rao T, Vijay Sundar Ram R, Vijayakrishna R and Sobha L. (2013). “A Text Chunker and Hybrid POS Tagger for Indian Languages.” In: Proceedings of the IJCAI, 2013.

Pattabhi R K Rao and Sobha Lalitha Devi. (2015) "Automatic Identification of Conceptual Structures Using Deep Boltzmann Machines" In Proceedings of Forum for Information Reterival and Evaluation-2015, ACM DL, ISBN 978-1-4503-4004-5

Pattabhi RK Rao and Sobha Lalitha Devi, (2017), ‘Patent Document Summarization Using Conceptual Graphs’, International Journal on Natural Language Computing (IJNLC), vol. 6(3).

Pattabhi RK Rao and Sobha Lalitha Devi, (2018), "Enhancing Multi-Document Summarization using Concepts", Sādhanā - Academy Proceedings in Engineering Sciences. vol 43:2 (27) . https://doi.org/10.1007/s12046-018-0789-y.

Ponnavaiko, M. 2002. “An Investigation on Unicode Standards for Tamil”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 29.

Prasad, V.K. Nagarajan, T. and Murthy H.A. (2004). “Automatic segmentation of continuous speech using minimum phase group delay functions.” Speech communication 42 (3-4), 429-446.

Prasanna Venkatesan, P and Chitrakekha T and Kuppuswami P. 2002. “PONN: A Tamil Operating Environment”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers.

Prathibha, P., Ramakrishana, A.G. 2002. “Web-enabled Speech Synthesizer for Tamil”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers, pages 134-140.

Prathibha, P., Ramakrishana, A.G. and Muralishankar, R. 2002. “Thirukural-II – A Text-to-speech Synthesis System”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers, 126-133.

Prathosh, A. P. 2013. T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," IEEE Transactions on Audio, Speech and Language Processing, 2013, Vol. 21, Iss. 12, pp. 2471-2480.

Pustejovsky, J. 1991. The generative lexicon. In Computational linguistics, vol 17, no 4, 409-441.

Pustejovsky, J. 1995. The Generative Lexicon. MIT Press, Cambridge, MA.

Rajan, K and Ganesan, M. 2002. “Tools for Corpus Analysis”, paper presented in International Conference on South Asian Linguistics (ICOSAL-4), 3-5 December, 2002, CASL, Annamalai University, Annamalainagar.

Rajendran. S. 1983. Semantics of Tamil Vocabulary. Report of the UGC sponsored postdoctoral work (manuscript). Poona: Deccan College Post-Doctoral Research Institute.

Rajendran, S. 1995. 'Towards a Compilation of a Thesaurus for Modern Tamil.' South Asian Language Review. 5.1:62-99.

Rajendran, S. 1995. “Machine bridge for Tamil and English." In: Proceedings of World Tamil Conference, Tamil Universiy, Thanjavur, 1995.

Rajendran, S. 1997. Grammatical Formalism and Computational Analysis of Nominal Compounds in Tamil. South Asian Language Review, vol. 7, no. 1, 1997, 27-46.

Rajendran, S. 1999. 'Spell and grammar checker for Tamil'. Paper read in 27th All India Conference of Dravidian Linguists held in ISDL, Thiruvananthapuram, 17th-19th, July, 1999.

Rajendran, S. 2002. Preliminaries to the preparation of Wordnet for Tamil. Language in India 2:1, March 2002, www.languageinindia.com.

Rajendran, S. 2003. Prerequisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages. Language in India 3:1, January 2003, www.languageinindia.com.

Rajendran, S. 2003. "Dravidian WordNet: a proposal", in R.M. Sundaram, et al (eds) Facets of Language, Thanjavur

Rajendran, S. 2003. Creating Generative Lexicon from MRDs: Tamil Experience. In: Rajeev Sangal et al (edited). Recent Advances in Natural Language Processing: Proceedings of the International Conference of Natural Language Processing (ICON 2003), 2003, 83-91.

Rajendran, S. 2003. "Creating Generative Lexicon form Dictionaries: Tamil Experience", in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Rajendran, S. 2003. "Prerequisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages", Language in India 3:1, www.languageinindia.com

Rajendran 2003. "Ontology and Knowledge Representation." Uploaded in academia.edu and Reseach Gate.

Rajendran, S. 2006. "Shallow Parsing in Tamil: the state of art", Language in India 6:7, www.languageinindia.com

Rajendran, S. 2006 "Morphological parsing on Tolkappiyam's perspective." Uploade in academia.edu and Reseach Gate.

Rajendran, S. 2006. Parsing in Tamil: Present state of art. Language in India 6:8, August 2006, www.languageinindia.com.

Rajendran, S. 2006. "Parsing in Tamil: Present state of Art. Indian Linguistics." vol. 67, 2006, 159–67.

Rajendran, S. 2006. “A Survey Of The State Of The State of Art In Tamil Language Technology.”. Language in India 6:10, October 2006, www.languageinindia.com

Rajendran, S. 2006. “Languge Technology in Tamil”, Language in India 6:8, www.languageinindia.com

Rajendran, S. 2007. ”Complexity of Tamil in POS Tagging.” Language in India 7:1, January 2007, www.langugeinindia.com.

Rajendran, S. 2009. ”Dravidian WordNet.” In: Proceedings of Tamil Internet Conference 2009. Cologne, Germany, October, 2009.

Rajendran, S. 2009. Dravidian WordNet. In: Proceedings of Tamil Internet Conference 2009. Cologne, Germany, October, 2009.

Rajendran, S. 2010. Tamil WordNet. In: Proceedings of the Global WordNet Conference (GWC 10) 2010, IIT, Bombay.

Rajendran S. 2014. Resolution of Lexical Ambiguity in Tamil. Language in India (e-journal), 14:1, January, 2014.

Rajendran S. 2018. “Lexical Resource Tool for Tamil Computing." In: Conferene Papers - 17th Tamil Internet Conference at Tamil Agricultural University, Coimpator.

Rajendran, S. 2012. “Preliminaries To The Preparation Of A Spell And Grammar Checker For TamiL” Upoaded in academia.edu and Reseach Gate.

Rajendran, S. 2016. “Tamil Thesaurus to Tamil wordNet.” In: Proceedings of 15th World Tamil Internet conference 2016, held at Gandhi Gram Rural University, Tamil Nadu, September 8-11, 2016, pages 1-9.

Rajendarn S. 2017. Technological Development of Tamil. International Journal of Central Univerisy of Kerala, Kasaragod. PP 73-118.

Rajendran, S. 2018. Visual Onto-Thesaurus for Tamil. In: Jayadevan V and Oppila Mathivanan (eds) Frontiers of Lexicography. Arimaa Nokku, Chennai, pp 101-113.

Rajendran S. 2018. Chapter 8: Computing Tools for Language Teaching and Learning. In: Muniisvaran, K. (ed.) Modern Studies in Language Teaching, pages 85-101.

Rajendran S. 2018 ”Lexical Resource Tools for Tamil Computing: History, Development and Future Prospects.” In: Conference papers, 17th Tamil Intenet Conference, Tamilnadu Agriculture University, Coimbatore.

Rajendarn S. 2017. Technological Development of Tamil. International Journal of Central University of Kerala, Kasaragod. PP 73-118.

Rajendran S. and Anandkumar M. 2017. “Computing Tools for Language Teaching and Language Learning.”. In: Conference papers, 17th Tamil Intenet Conference, Tamilnadu Agriculture University, Coimbatore.

Rajendran, S., M. Anandkumar and Soman K.P. 2013. Computational approach to Word Sense disambiguation in Tamil. In: 12th International Tamil Internet Conference 2013, University of Malaya, Kuala Lumpur, Mayasia.

Rajendran S and Anandkumrar M. 2014. “Corpus based approach for resolving verbal polysemy in Tamil”. In: Proceedings of 13th International Conference on Tamil Computing and Tamil Internet (Tamil Internet 2014) from 19th to 21st September 2014 at Pondicherry University.

Rajendran S and Anandkumrar M. 2014. “Corpus based approach for resolving verbal polysemy in Tamil”. In: Proceedings of 13th International Conference on Tamil Computing and Tamil Internet (Tamil Internet 2014) from 19th to 21st September 2014 at Pondicherry University.

Rajendran, S. Anand Kumar M. and Soman K.P. (2015). “Building hierarchies and networks from MRDs of Tamil”. In: Proceedings of 14th International Conference on Tamil Computing and Tamil Internet (Tamil Internet 2015), at Singapore (SIM University, Singapore campus) from 30th, 31st May to 1st June 2015.

Rajendran, S. and Anandkumar, M. 2017. “Visual Onto-Thesaurus for Tamil.” Language in India www.languageinindia.com ISSN 1930-2940 vol. 17:5 May 2017.

Rajendran S and Anand Kumar 2018. “Computing tools for Tamil Language Teaching and Learning.” In: Conferene Papers - 17th Tamil Internet Conference at Tamil Agriculural University, Coimpator.

Rajendran, S., Arulmozi S., Kumara Shanmugam B., Baskaran S. & Thiagarajan S. 2002. Tamil WordNet. In: Proceedings of the First Global WordNet Conference, CIIL, Mysore, 2002, 271-274.

Rajendran S, Arulmozi S, Ramesh Kumar, Viswanathan S. 2003. "Computational morphology of verbal complex," B. Ramakrishna Reddy (ed) Word Structure in Dravidian, Dravidian University, Kuppam, 2003, 376-398.

Rajendran, s. and Baskaran, S. 2002. "Preparation of Electronic Thesaurus for Tamil", in Proceedings of the International Conference on Natural Language Processing. Mumbai: NCST.

Rajendran, S. and Kamakshi, S. 2002. "Preliminaries to the preparation of a machine translation aid to translate Linguistics Texts in English into Tamil", Paper read in UGC-SAP National Seminar 'On Translation' 7th-9th March 2002, CAS in Linguistics, Annamalai University.

Rajendran S. and S. Kamakshi. 2003. Preliminaries to the Preparation of a Machine Aid to Translate Linguistic Text Books in English into Tamil.. Language in India 3:9, September, 2003, www.languageinindia.com.

Rajendran S. and S. Kamakshi. 2008. Preliminaries to the Preparation of a Machine Aid to Translate Linguistic Texts in English into Tamil. DLA Publications, Thivuvananthapuram.

Rajendran, S, Shivapratap G, Dhanalakshmi V and Soman K.P. 2010. Building a WordNet for Dravidian Languages. In: Proceedings of the Global WordNet Conference (GWC 10), 2010 IIT Bombay.

Rajendran, S. and Mohan Raj, S.N. 2016. Tamil Oriented Machine Translation System under Indian Language to Indian Language Machine Tranlsation (ILILMT) Consortium. In: 15th World Tamil Internet Conference-2016: Conference Papers. Gandigram University, Didikal.393-402.

Rajendran S and Vasuki G. 2013. English To Tamil Machine Translation System Using Parallel Corpus. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:5 May 2019 . (Book Uploaded in academia.edu and Research Gate.)

Rajeswari Sridhar, Rasiga Gowrisankar and G. Monica. 2017. "Language Relationship Model for Automatic Generation of Tamil Stories from Hints." International Journal of Intelligent Information Technologies (IJIIT) 13(2).

Ramakrishnan, A.G. 2001 “Issues in Standardization for Text to Speech in Tamil”, Tamilnet 2001, Kulalumpur, Malasia.

Ramakrishnan, A.G. 2014. Speech Technology and Tamil. Conference Paper (PDF Available). January 2014 DOI: 10.13140/RG.2.1.2287.4725 Conference: National Conference on Tamil Internet, At Chennai, India.

Ramakrishnan, A.G. Lakshmish N Kaushik, Laxmi Narayana M. 2007. Natural Language Processing for Tamil TTS. Conference: 3rd Language and Technology Conference (LTC), At Poznan, Poland October 2007.

Ramakrishnan A G, B Abhiram and S R Mahadeva Prasanna. 2015. “Voice source characterization using pitch synchronous discrete cosine transform for speaker identification,” Journal of the Acoustical Society of America Express Letters, Vol. 137(), pp., 2015.

Ramaswamy, V. 2000. Morphological Generator for Tamil. Unpublished M.Phil. dissertation. University of Hyderabad.

Ramaswamy, Vaishnavi. 2003. A Morphological Analyzer for Tamil. Unpublished Ph.D Dissertaion Subbmitted to University of Hyderabad.

Ramasamy, L. 2011. "TamilTB: An Effort Towards Building a Dependency Treebank for Tamil." WDS'11 Proceedings of Contributed Papers, Part I, 143–148, 2011. ISBN 978-80-7378-184-2 © MATFYZPRESS

Ramasamy, L. and Žabokrtský Z. 2011. Tamil Dependency Parsing: Results Using Rule Based and Corpus Based Approaches. International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2011: Computational Linguistics and Intelligent Text Processing pp 82-95.

Ramaswamy, Vaishnavi. 2000. A Morphological Generator for Tamil. Unpublished M.Phil Dissertation. Hyderabad: University of Hyderabad.

Ranganathan, Vasu. 1997. ‘A Lexical Phonological Approach to Tamil Word by computer’. International Journal of Dravidian Linguistics 26.1:57-70.

Ranganathan, V. 1997. “A Lexical Phonology Approach to Tamil Words by Computer”, International Journal of Dravidian Linguistics 26:1.57-70.

Ranganathan, Vasu. 2001. “Development of Part-of-Speech Tagger for Tamil.” Tamil Internet Conference, 2001.
(http://infitt.org/drupal17a/ TIconferencepapers/TI2001/vasur.pdf).

Ranganathan, Vasu. (2002). “An Interactive approach to development of English to Tamil Translation machine Translation System on the web.” Tamil Internet Conference, 2002. (http://infitt.org/drupal17a/TIconferencepapers/TI2002/15VASUR.PDF).

Ranganathan, Vasu. Computational Phonology and the Development of Text-to-Speech Application for Tamil.

Ranganathan, Vasu. 2001. "Development of Part-of-Speech Tagger for Tamil", Tamil Internet Conference, 2001, (http://infitt.org/drupal7a/TIconferencepapers/TI2001/vasur.pdf).

Ranganathan, Vasu (2002). "An interactive approach to development of English to Tamil Translation machine Translation system on the web", Tamil Internet Conference, 2002.
(http://infitt.org/drupal7a/TIconferencepapers/TI2002/15VASUR.PDF)

Ranganathan, Vasu. 2010. Tamil Language in Context: A Comprehensive Approach to Learning Tamil, Department of South Asia Studies, University of Pennsylvania, http://www.thetamillangauge.com.

Renganathan, Vasu. 2014. “Computational Phonology and the Development of Text-to-Speech Application for Tamil.” Tamil Internet Conference, 2014, Pondicherry University: Pondicheery.

Ranganathan, Vasu. 2016a. Chapter 2: Computational Phonology. In: Computational Approaches To Tamil Linguistics. Chennai: Cre-A Publications.

Ranganathan, Vasu. 2016b. Chapter 3: Computational Morphology. In: Computational Approaches To Tamil Linguistics. Chennai: Cre-A Publications.

Ranganathan, Vasu. 2016c. Chapter 4: Computational Syntax. In: Computational Approaches To Tamil Linguistics. Chennai: Cre-A Publications.

Ranganathan, Vasu. 2016d. Chapter 5: Computational Semantics. In: Computational Approaches To Tamil Linguistics. Chennai: Cre-A Publications.

Ranganathan, Vasu. 2016e. Chapter 6: Applications of Tamil NLP Systems. In: Computational Approaches To Tamil Linguistics. Chennai: Cre-A Publications.

Ranganathan, Vasu. 2016f. Computational Approaches To Tamil Linguistics. Chennai: Cre-A Publications.

Rekha R U, Anand Kumar M, Dhanalaksmi V, Soman K P and Rajendran S. A Novel Approach to Morphological Generator for Tamil. Data Engineering and Management, Second International Conference, ICDME, Springer-Verlag, Berlin, Heidelberg, 2012, 249-251.

Remmiya Devi G, Veena P V, Anand Kumar M and Soman K P AMRITA_CEN@FIRE 2016: Code-Mix Entity Extraction for Hindi-English and Tamil-English Tweets.

Richie D.G., Russel J.G. Black W.A., and Pulman G.S. 1992. Computational Morphology. Practical mechanisms for English Lexicon. Massachusetts: Massachusetts Institute of Technology.

Samuel Thomas, Venugopalakrishna, Y R. and Murthy H.A. (2007). “Speech Synthesis using Syllable-like units,” Workshop on Speech Synthesis, CDAC, Noida, April 13-14, 2007.

Sanjay S, Anand Kumar M., and Soman K. P. (2015) AMRITA-CEN-NLP@FIRE 2015:CRF based named entity extraction for Twitter microposts. CEUR Workshop Proceedings, 1587:96–99, 2015.

Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharya, Girish Nath Jha, Rajendran S, Saravanan. K, Sobha L and Subbarao, K. V. (2008). A Common Part of Speech Tagset Framework for Indian Languages, LREC 2008, 6th Language Resources and Evaluation Conference, Marrakech, Morocco.Brouchers on ‘Language Technology Products’ of the Resource Center for Indian Language Technology Solutions – Tamil, Chennai

Sankaralingam C, S. Rajendran, B. Kavirajan, M. Anand Kumar, K. P. Soman. 2017. “Onto-thesaurus for tamil language: Ontology based intelligent system for information retrieval.” ICACCI 2017: 2396

Sarada G.L., Lakshmi A, Murthy H.A. and Nagarajan T. (2009).“Automatic Transcription of Continuous Speech into Syllable-like units for Indian Languages,” Sadhana, Indian Academy of Sciences,Vol. 34, Part 2, April 2009, pp. 221-233.

Saraswathi S. and Geetha T. V. (2007), ‘Comparison of Performance of Enhanced Morpheme-based language Model with Different Word-based Language Models for Improving the

Performance of Tamil Speech Recognition System', ACM Transaction on Asian Language Information Processing Vol. 6 No. 3, Article 9.

Saraswathi,S., Kanivadhana, P., Anusiya, M., & Sathiya, S. (2011). Bilingual Translation System (For English and Tamil), International Journal on Computer Science and Engineering (IJCSE),3(3).Saravanan, S., Menon, A. G., &Soman, K. P. (2010). English to Tamil Machine Translation System, Proceedings of INFITT-2010, at Coimbatore.

Schiffman, H.F. and Vasu Rengnatha, "Tamil Script Reform and Glyph Rendering Approach in Unicode: Past and Present Attempts to Simplify Tamil Writing System" In: Peri Bhaskararao (ed.). Working Papers of International Symposium on Indic Scripts: Past and Present. , ILCAA:Tokyo. Selvam M, Natarajan. A M, and Thangarajan R. (2008). Structural Parsing of Natural Language Text in Tamil Using Phrase Structure Hybrid Language Model. World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:2, No:3, 2008

Saint-Dizer P and Viegas, E. 1995. "An Introduction to Lexical semantics from a linguistic and psycholinguistic perspective". In: Saint-Dizer P and Viegas, E (eds) Computational Lexical Semantics. Cambridge: Cambridge University Press.

Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy, C. S. Ramalingam. 2006. "Natural Sounding TTS based on Syllable-like Units, Proceedings of 14th European Signal Processing Conference, Florence, Italy, Sep 2006.

Santiago Fernandez, Alex Graves, and Jürgen Schmidhuber. 2007. An application of recurrent neural networks to discriminative keyword spotting. Proceedings of ICANN (2), pp. 220–229.

Sarkar, A. 1993. Extending Kimmo's two-leveal model of morphology. CDAC Pune: Pune Univerisity.

Sarada, G. L. , N. Hemalatha, T. Nagarajan, and Hema A.Murthy. 2004. "Automatic transcription of continuous speech using unsupervised and incremental training", INTERSPEECH - 2004, pp 405-408, Korea.

Sarada, G. L., Nagarajan, T., Hema A. Murthy. 2004 "Multiple Frame Size And Multiple Frame Rate Feature Extraction For Speech Recognition", SPCOM-2004, December.

Saraswathi. S and Geetha T. V., 2004. Implementation of Tamil Speech Recognition System Using Neural Networks, Lecture Notes in Computer Science, Vol. 3285.

Saraswathi. S and Geetha T. V., 2007. Comparison of Performance of Enhanced Morpheme-based language Model with Different Word-based Language Models for Improving the Performance of Tamil Speech Recognition System, ACM Transaction on Asian Language Information Processing Vol. 6 No. 3, Article 9.

Sarveswaran, Kengatharaiyer and Mahesan, Sinnathamby. 2014. Hierarchical Tag-set for Rule-based Processing of Tamil Language. International Journal of Multidisciplinary Studies (IJMS) Volume I, Issue I, 2014.

Schmidhuber, Jürgen. 2015. "Deep learning in neural networks: An overview". Neural Networks. 61: 85–117. arXiv:1404.7828. doi:10.1016/j.neunet.2014.09.003. PMID 25462637.

Selvam, M. Natarajan, A. M. and Thangarajan R. 2009. "Structural Parsing of Natural Language Text in Tamil Language Using Dependency Model", International Journal of Computer Processing of Languages Vol. 22, No. 02n03, pp. 237-256 (2009) https://doi.org/10.1142/S1793840609002093.

Selvam M, Natarajan. A M, and Thangarajan R .2008. "Structural Parsing of Natural Language Text in Tamil Using Phrase Structure Hybrid Language Model" , International Journal of Computer and Information Engineering.

Selvam M., A.M. Natarajan. 2009. Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques. International Journal Of Computers, Issue 4, Volume 3, 2009.Shanmugam, C. 2001. “Computer Analysis of Simple Sentence in Tamil”. Paper read in UGC-SAP National Seminar on Computational Linguistics and Dravidian Languages, 22-24 February, 2001, CAS in Linguistics, Annamalai University, Annamalainagar.

Sengupta Probal and Chauduri B B .1997. 'A Delayed Syntactic-Encoding-based LFG Parsing Strategy for an Indian Language-Bangla', ACL, Madrid, Spain.

Sepp Hochreiter; J. Schmidhuber .1997. "Long Short-Term Memory". Neural Computation. 9 (8): 1735-1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276.

Shanmugan, C. 2001. “Computer Analysis of Simple Sentences in Tamil”, Paper read in UGC-SAP National Seminar on Computational Linguistics and Dravidian Languages, 22-24 February, 2001, CAS in Linguistics, Annamalai University, Annamalainagar

Shanmugan, C. 2002. ”Grammar and Parser: A Program for Syntactic Parsing in Tamil”, International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Shanmugan, C. “Minimalist Program for Tamil Parsing”, Seminar paper.

Shanmugam, C. 2002. “Grammar and Parser: A Program for Syntactic Parsing in Tamil,” International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Shanmugam, C. “Minimalist Program for Tamil Parsing”.Shriya, S. Vinayakumar R, Anand Kumar M, and Soman K. P. (2015) AMRITA-CEN@ SAIL2015: Sentiment Analysis in Indian Languages. In International Conference on Mining Intelligence and Knowledge Exploration, pages 703–710. Springer, 2015.

Shivkaran Singh, M. Anand Kumar, K. P. Soman. 2018. Attention based English to Punjabi neural machine translation. Journal of Intelligent and Fuzzy Systems 34(3): 1551-1559 (2018)

Sivashanmugam, C. (2000) “A Model for Computer analysis of Verbs in Tamil”. in Working Papers in Linguistics, Department of Linguistics, Bharathiyar University, Coimbatore.

Sivashanmugam, C. 2002. Morphological Processor for Negative Constructions in Tamil”. Indian Conference on Natural Language Processing, Anna University, Chennai.

Shriya Se, R. Vinayakumar, M. Anand Kumar and K. P. Soman. 2016. "Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms." Indian Journal of Science and Technology, Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/106482, December 2016.

Sindhuja Gopalan, Bakyavathi T, Vijay Sundar Ram, Sobha L. (2012)."Analysis of 'genetive drop' in Tamil-Hindi machine translation", presented in 40th All India conference for Dravidian Linguistics held at CALTS, School Of Humanities, University Of Hyderabad.

Singh R. and Ford A. 1984. “Interference, overgeneralization, adaptive rules and phonological theory”. Occasional papers in Linguistics, Carbonale: University of Illinois Press.

Sobha, L. 2003. "Pronominal Resolution in South Dravidian languages", 23rd South Asian Language Analysis, University of Texas, Austin, 10-12 October 2003.

Sobha, L and Vijay Sundar Ram. 2006. “Noun Phrase Chunker for Tamil Language”, in Proceedings of the First National Symposium on Modeling and Shallow Paring of Indian Languages, pages 194-198.

Sobha L and Vijay Sundar Ram R. 2007. "Multilingual Place Name Tagger for Indian Languages", In the Proceedings of IJCAI Workshop on Cross Lingual Information Access (CLIA) - Addressing the Information Need of Multilingual Societies, Hyderabad. pp 34 – 39

Sobha Lalitha Devi, Sindhuja Gopalan, R. Vijay Sundar Ram., 2013. ”Transfer Grammar in Tamil-Hindi MT System”, In Proceedings of International Conference on Asian Language Processing 2013, Urumqi, China. pp 79-82.

Sobha Lalitha Devi, Vijay Sundar Ram, Pattabhi RK Rao. 2014. "A Generic Anaphora Resolution Engine for Indian Languages". In Proceedings of the 25th International Conference on Computational Linguistics (Coling 2014).

Sobha Lalitha Devi, Vijay Sundar Ram R., and Pattabhi RK Rao. 2014. “Anaphora Resolution System for Indian Languages”, In proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation (WILDRE), Organized under LREC2014, Reykjavik, Iceland

Sobha Lalitha Devi, Sindhuja Gopalan, Lakshmi S. 2014. “Automatic Identification of Discourse Relations in Indian Languages”, In proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation, Organized under LREC2014, Reykjavik, Iceland.

Sobha L, Vijay Sundar Ram R. (2006). “Noun Phrase Chunking in Tamil.” In: Proceedings of the Sympositum on Modeling and Shallow Parsing of Indian Languages, Indian Institute of Technology, Bombay. pp:194-198.

Sobha Lalitha Devi, Vijay Sundar Ram, Pattabhi RK Rao. 2014. "A Generic Anaphora Resolution Engine for Indian Languages". In Proceedings of the 25th International Conference on Computational Linguistics (Coling 2014).

Sobha Lalitha Devi, Vijay Sundar Ram R., and Pattabhi RK Rao 2014. “Anaphora Resolution System for Indian Languages”, In proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation (WILDRE), Organized under LREC2014, Reykjavik, Iceland

Sobha Lalitha Devi, Sindhuja Gopalan, Lakshmi S. 2014. “Automatic Identification of Discourse Relations in Indian Languages”, In proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation, Organized under LREC2014, Reykjavik, Iceland.

Sobha Lalitha Devi, Lakshmi S and Sindhuja Gopalan. 2014. “Discourse Tagging for Indian Languages”, In A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol 8403, pp. 469-480.

Sobha Lalitha Devi, Lakshmi S and Sindhuja Gopalan. 2014. “Discourse Tagging for Indian Languages”, In A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol 8403, pp. 469-480.

Sobha Lalitha Devi and Pattabhi RK Rao.2016. "Mining of Social Networks from Literary Texts of Resource Poor Languages", In proceedings of CICLING 2016.In proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey, Springer LNCS

Sobha Lalitha Devi and Pattabhi RK Rao.(2016). "Mining of Social Networks from Literary Texts of Resource Poor Languages", In proceedings of CICLING 2016.In proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey, Springer LNCS.

Sobha Lalitha Devi and Pattabhi RK Rao.2016. "Mining of Social Networks from Literary Texts of Resource Poor Languages", In proceedings of CICLING 2016.In proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey, Springer LNCS

Sobha Lalitha Devi and Pattabhi RK Rao. 2016. "Semantic Representation of Tamil Texts using Conceptual Graphs", In Proceedings of 3rd Workshop on Indian Language Data: Resources & Evaluation (WILDRE-3) under LREC 2016 in Portoroz, Slovenia

Sobha Lalitha Devi, Vijay Sundar Ram, Sindhuja Gopalan, Pattabhi RK Rao and Lakshmi S. 2016. "A Demo Proposal: Tamil – Hindi Automatic Machine Translation – A detailed Description", In Proceedings of 3rd Workshop on Indian Language Data: Resources & Evaluation (WILDRE-3) under LREC 2016 in Portoroz, Slovenia

Sobha Lalitha Devi, Pattabhi RK Rao, Vijay Sundar Ram and Malarkodi C.S.2016. "A Demo Proposal: Tamil – English Cross Lingual Information Access (CLIA) System", In Proceedings of 3rd Workshop on Indian Language Data: Resources & Evaluation (WILDRE-3) under LREC 2016 in Portoroz, Slovenia.

Sobha Lalitha Devi and Pattabhi RK Rao.(2016). "Semantic Representation of Tamil Texts using Conceptual Graphs", In Proceedings of 3rd Workshop on Indian Language Data: Resources & Evaluation (WILDRE-3) under LREC 2016 in Portoroz, Slovenia

Sobha Lalitha Devi, Vijay Sundar Ram, Sindhuja Gopalan, Pattabhi RK Rao and Lakshmi S. 2016. "A Demo Proposal: Tamil – Hindi Automatic Machine Translation – A detailed Description", In Proceedings of 3rd Workshop on Indian Language Data: Resources & Evaluation (WILDRE-3) under LREC 2016 in Portoroz, Slovenia

Sobha Lalitha Devi, Pattabhi RK Rao, Vijay Sundar Ram and Malarkodi C.S.2016. "A Demo Proposal: Tamil – English Cross Lingual Information Access (CLIA) System", In Proceedings of 3rd Workshop on Indian Language Data: Resources & Evaluation (WILDRE-3) under LREC 2016 in Portoroz, Slovenia.

Sowa, J.F. 1984. Conceptual Structures: Information Processing in Mind and Machine. Reading, Massachusetts: Addison-Wesley Publishing Company.

Spark Jones, K. 1986. Synonymy. Edinburgh University Press.

Spencer, A. 1991. Morphological Theory. An introduction to word structure in generative grammar. Oxford: Basil Blackwell.

Spencer, A. and Zwicky, A. (eds) 1998. Hand book of morphology. Oxford: Blackwell.

Sproat, R. 1992. Morphology and Computation. A Bradford Book, The MIT Press, Cambridge.

Stevens, Kenneth N. 2000. Acoustic Phonetics. Current Studies in Linguistics (No. 30). Cambridge, MA: MIT. ISBN 0-262-69250-3.

Stevens, Kenneth N. 2002. "Toward a model for lexical access based on acoustic landmarks and distinctive features". The Journal of the Acoustical Society of America. 111 (4): 1872–1891. doi:10.1121/1.1458026. PMID 12002871.

Stump, Gregory T. 1998. “Morphological operations”. In Andrew Spencer and Arnold M. Zwicky (eds.), The Hand book of morphology.123-143, Oxford:Blackwell.

Stump, Gregory T. 2001. Inflectional Morphology: A Theory of Paradigm Structure (Cambridge Studies in Linguistics 93). Cambridge University Press, hardback ISBN: 0-19-512600-9, xvi+308pp.

Subramoniam V. I., Mahidas Bhattacharya, A. Lohy, S. Tarai. 2001. Speech Synthesis (Tamil & Oriya): an application for the Blind, (III.5(35)/2001-ET), Department of Science and Technology, Govt. of India.

Sundaram, L. 2015.Tamil Computing, Vikatan Publications, Chennai, May 2015.

Sureka K, Srinivasagan K.G. and Suganth S. An Efficient Dependency Parser Using Hybrid Approach For Tamil Language.

Suresh, R.M., Arumugan, S., and Aravanan, K.P. 2000. “Recognition of Handwritten Tamil characters using fuzzy classificatory approach”, in Proceedings of the Tamil Internet 2000 Conference, Singapore.

Taft, M., & Forster, K. I. 1975. Lexical storage and retrieval of prefixed words. Journal of Verbal Learning and Verbal Behavior, 14(6), 638-647.

Sureka K, Srinivasagan K.G. and Suganth S. An Efficient Dependency Parser Using Hybrid Approach for Tamil Language.

Thayaparan M., Ranathunga S. and Thayasivam U. 2018. "Graph Based Semi-Supervised Learning for Tamil POS Tagging", LREC 2018.

Thangarajan R., Natarajan A.M. and Selvam M., 2008. Word and Triphone based Approaches in Continuous Speech Recognition for Tamil Language, WSEAS Transactions on Signal Processing, pp. 76- 85.

Thangarajan R. and A.M. Natarajan. 2008. ”Syllable Based Continuous Speech Recognition for Tamil.” South Asian Language Review, VOL.XVIII. No. 1, January 2008.

Thangarajan R. Speech Recognition for Agglutinative Languages. http://dx.doi.org/10.5772/50140

Thangarajan R., Natarajan A. M. and Selvam M. 2008a. ‘Word and Triphone based Approaches in Continuous Speech Recognition for Tamil Language’, WSEAS Transactions on Signal Processing, Issue 3, Vol.4, 2008, pp. 76-85.

Thangarajan R. and Natarajan A. M. 2008b. 'Syllable Based Continuous Speech Recognition for Tamil Language', South Asian Language Review (SALR), Vol. XVIII, No. 1, pp. 71-85.

Thangarasu M. and R.Manavalan. 2013. "Stemmers for Tamil Language: Performance Analysis," International Journal of Computer Science & Engineering Technology (IJCSET)ISSN : 2229-3345 Vol. 4 No. 07 Jul 2013, 902-908.

'The Unicode Standard: A Technical Introduction' 8/05/2001

Theivendiram P, Megala Uthayakumar, Nilusija Nadarasamoorthy, Mokanarangan Thayaparan, Sanath Jayasena, Gihan Dias Surangika Ranathunga. 2016. "Named-Entity-Recognition (NER) for Tamil Language Using Margin-Infused Relaxed Algorithm (MIRA)." International Conference on Intelligent Text Processing and Computational Linguistics, Pages 465-476, Springer, Cham.

Thangarajan R. Speech Recognition for Agglutinative Languages. http://dx.doi.org/10.5772/50140

Thangarajan R. and Natarajan A. M. 2008b. 'Syllable Based Continuous Speech Recognition for Tamil Language', South Asian Language Review (SALR), Vol. XVIII, No. 1, pp. 71-85.

Thangarajan R., Natarajan A. M. and Selvam M. 2008a. 'Word and Triphone based Approaches in Continuous Speech Recognition for Tamil Language', WSEAS Transactions on Signal Processing, Issue 3, Vol.4, 2008, pp. 76-85.

Thennarasu Skkan. 2016. "Tamil Pronouncing Dictionary: A Promising Trend in Language Technology Development." Journal of Advanced Linguistic Studies, Vol. 5, No. 1-2, Jan-Dec 2016

Thilagavathi R and Krishnakumari K. 2016. "Tamil English Languge Sentiment Analysis." International Journal of Engineering Research and Technology (IJERT) ICETET-2016.

Tyler, A., & Nagy, W. 1990. Use of derivational morphology during reading. Cognition, 36, 17-34.

Udhyakumar N., C.S.Kumar, R.Srinivasan and R.Swaminathan. 2004." Decision Tree Learning for Automatic Grapheme to Phoneme Conversion for Tamil. Down loaded form internet.

Vaishnavi Ramaswamy. 2000. A Morphological Generator for Tamil. M.Phil Dissertaion Subbmitted to University of Hyderabad.

Vaishnavi Ramaswamy. 2003. A Morphological Analyzer for Tamil. Ph.D Dissertaion Subbmitted to University of Hyderabad.

Vaishnavi, Ramaswamy. 2003. “Parsing in AMPLE, KIMMO & PERL: Nouns in Tamil”, Paper presented in

Vaishnavi, T. and Roxanna Samuel. 2016. "Individual Document Keyword Extraction for Tamil". International Journal of Computer Technology & Applications,Vol 7(3), 448-452 IJCTA May-June 2016 Available online@www.ijcta.com

Vaishnavi, T. and Roxanna Samuel. 2016. "Individual Document Keyword Extraction for Tamil". International Journal of Computer Technology & Applications,Vol 7(3), 448-452 IJCTA May-June 2016 Available online@www.ijcta.com.

Vasudevan, V. 2002. “Character Recognition Techniques–A Demo Program”, International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Vasu, R. 1993. “A Logical approach to development of natural language understanding system for Tamil.” PJDS 3:1, 53-64.

Vasu, R. 1997a “A Lexical Phonological Approach to Processing Tamil Word by Computer.” IJDL, 26:1, January 1997.

Vasu, R. 1997b. “Significance of Creation and Use of Corpus of Modern Tamil Prose Text through the Web.” International Symposium for Tamil Information Processing and Resources on the Internet, National University of Singapore, May 1997.

Vasu, R. 1988. “Human Aided Machine Translation (Problems and Perspectives). In: Karunakaran and Jayakumar (eds.)

Vasu Renganathan. 2002. “Interactive Approach to Development of English-Tamil Machine Translation System on Web”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 68-73.

Vasudevan, V. 2002. “Character Recognition Techniques–A Demo Program”, International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Vendler, Z. 1967. Linguistics and Philosophy. Ithacha, NY: Cornell University Press.

Veera Alagiri, S. 2011. Tamil to Malayalam Machine Translation System. Unpublished PhD dissertation, Tamil University, Thanjavur.

Vignesh, N & S. Sowmya Automatic Question Generator in Tamil. International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 10, October - 2013Vijayakrishna R and Sobha L, (2008), "Domain focused Named Entity Recognizer for Tamil using Conditional Random Fields", NERSSEAL 08, IJCNLP2008 workshop,

Vignesh N and S.Sowmya. “Automatic Question Generator in Tamil.” International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 10, October - 2013 IJERTIJERT ISSN: 2278-0181

Vijayakrishna R and Sobha L. Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields.

Vijayakrishna R and Sobha L. 2008. Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields. in Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages, Hyderabad, India, 2008, pp. 59-66.

Vijay Sundar Ram R., Chandra Mouli N., Bhuvaneswari P., Ananda Priya J. and Kumara Shanmugam B. 2005. “Hybrid Approach for Developing a Tamil Spell Checker.” In the Proceedings of International Conference on Natural Language, Indian Institute of Technology, Kanpur, pp. 111-115.

Vijay Sundar Ram and Sobha L, 2007. "Automatic Identification of Semantic Arguments" Fifth Malaysia International Conference on Languages, Literatures, and Cultures (MICOLLAC '07), Malaysia.

Vijay Sundar Ram R. and Sobha L.. 2008. "Clause Boundary Identification Using Conditional Random Fields", In Proceedings of 9th International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel.

Vijay Sundar Ram R, Menaka S and Sobha Lalitha Devi. 2010. “Tamil Morphological Analyser”, in “Morphological Analysers and Generators”, (ed.) Mona Parakh, LDC-IL, Mysore, pp. 1-18.

Vijay Sundar Ram R. and Sobha Lalitha Devi. 2010. "Noun Phrase Chunker Using Finite State Automata for an Agglutinative Language", Proceedings of the Tamil Internet - 2010 at Coimbatore, India, June 23-27, pp. 218- 224

Vijay Sundar Ram R, Bakiyavathi T, Sindhujagopalan R, Amudha K and Sobha Lalitha Devi. 2012. "Tamil Clause Boundary Identification: Annotation and Evaluation" In Proceedings of Workshop on Indian Language Data: Resources and Evaluation, LREC 2012, Istanbul

Vijay Sundar Ram and Sobha Lalitha Devi. 2012. "Coreference Resolution using Tree-CRF", In A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol. 7181/2012 pp 285 - 296.

Vijay Sundar Ram and Sobha Lalitha Devi. 2012. "Coreference Resolution using Tree-CRF", In A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol. 7181/2012 pp 285 - 296.

Vijay Sundar Ram, R. and Sobha Lalitha Devi. 2013. “Pronominal Resolution in Tamil Using Tree CRFs”, In Proceedings of International Conference on Asian Language Processing 2013, Urumqi, China. pp197-200.

Vijay Sundar Ram, R. and Sobha Lalitha Devi. 2013. ”Pronominal Resolution in Tamil Using Tree CRFs”, In Proceedings of 6th Language and Technology Conference, Human Language Technologies as a challenge for Computer Science and Linguistics -2013, Poznan, Poland

Vijay Sundar Ram and Sobha Lalitha Devi, 2016. Two Layer Machine Learning Approach for Mining Referential Entities for a Morphologically Rich Language. Asian Journal of Information Technology, 15: 2831-2838.

Vijay Sundar Ram and Sobha Lalitha Devi. 2016. "How to Handle Split Antecedents in Tamil?", In proceedings of Coreference Resolution Beyond OntoNotes co-located with NAACL 2016, San Diego, California.

Vijay Sundar Ram and Sobha Lalitha Devi. 2016. Two Layer Machine Learning Approach for Mining Referential Entities for a Morphologically Rich Language. Asian Journal of Information Technology, 15: 2831-2838.

Vijay Sundar Ram and Sobha Lalitha Devi. 2016. "How to Handle Split Antecedents in Tamil?", In proceedings of Coreference Resolution Beyond OntoNotes co-located with NAACL 2016, San Diego, California.

Vijay Sundar Ram and Sobha Lalitha Devi. 2017. "A Robust Coreference chain Builder for Tamil", In proceedings of 18th International Conference on Computational Linguistics and Intelligent Text Processing, April 17 to 23, 2017, Budapest, Hungary

Vijay Sundar Ram and Sobha Lalitha Devi. 2018. "A Semi-automated Annotation of Co-reference Chains in Tamil", In proceedings of 18th International Conference on Computational Linguistics and Intelligent Text Processing, March 18 to 24, 2018, Hanoi, Vietnam,

Viswanathan, S., Rameshkumar, S. Kumara Shanmugam, B. Arulmozhi. S. & Vijay Shnakar, K.2003. “A Tamil Morphological Analyzer”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL.

Viswanathan S. 2000. Tamil Morphological Analyser. MS Thesis. Anna University, Chennai.

Viswanathan, S. Ramesh Kumar, S. Kumara Shanmugham, B. Arulmozi, S. and Vijay Shanker K. 2003. A Tamil Morphological Analyser, ICON-2003, 31-39.

Wagner R.A. and Fischer M.J. 1974. ”The string to string correction problem”. Journal of the Association for Computing Machinery, 21, 168-173.

Weerasinghe, R. (2011). A SMT Approach to Sinhala-Tamil Language Translation. citeseerx.ist.psu.edu /viewdoc/summary?doi= 10.1.1.78.7481, 2011.

Winston Cruz, S. 2002. Parsing and Generation of Tamil Verbs in GSMorph. M.Phil. dissertation submitted to the University of Hyderabad.

Weerasinghe, R. 2011. A SMT Approach to Sinhala-Tamil Language Translation. citeseerx.ist.psu.edu /viewdoc/summary?doi= 10.1.1.78.7481, 2011.

Weizenbaum, J. 1966. “ELIZA – A computer program for the study of natural language communication between man and machine”. Communications of the ACM, 9 (1), 36-45.

Weber J.D., Black H.A. McConnel R.S. 1998. AMPLE. A tool for exploring morphology. Dallas: Summer Institute of Linguistics.

Wilson, M and Emmorey K . 1998. A visuospatial “phonological loop” in working memory: Evidence from American Sign Language. Memory & Cognition 25 (3), 313-320.

Winiwarter Werner.1995. 'Extended CUG for Free Word Order Languages and its Efficient Implementation within an IDA Architecture', 8th Asian Conference on Language, Information and Computation.

Winston, M.E. and Chaffin R and Hermann D 1987. A taxonomy of part-whole relations. Cognitive science, 11, 417-444.

Winston Cruz, S. 2002. Parsing and Generation of Tamil Verbs in GSMorph. M.Phil. dissertation submitted to the University of Hyderabad.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proceedings of the 14th International Conference on Computational Linguistics (COLING, Nantes, France). 454–460.

Yarowsky, D. 1993. One sense per collocation. In Proceedings of the ARPA Workshop on Human Language Technology (Princeton, NJ). 266–271.

Yarowsky, D. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Las Cruces, NM). 88–95.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (Cambridge, MA). 189–196.

Yarowsky, D. 2000. Hierarchical decision lists for word sense disambiguation. Comput. Human. 34, 1-2, 179–186.

Yarowsky, D. and Florian, R. 2002. Evaluating sense disambiguation across diverse parameter spaces. J. Nat. Lang. Eng. 9, 4, 293–310.

Yagnanarayana, B. et al. 1992. "Text-to-speech system for Indian Languages", presented at the Workshop on Computer Applications in Indian Languages held at CIIL, Mysore, August 19-21

Yegnanarayanan B, Rajendran, S, Ramachandaran V.R., Madhukumar, A.S. "Significance of Knowledge Sources for a Text-to-speech System for Indian Languages", in Sadhana, Academic Proceedings in Engineering Sciences of India.

Yegnanarayanan, B. 1994. "Speech synthesis by Machine", in B.B. Rajpurohit (ed.) Technology and Languages. Mysore:CIIL.

Yegnanarayana.B and Nayeemullah Khan.A 2001. Development of a speech recognition system for Tamil for restricted small tasks. In Proc. NCC, India.